

## Simulations of the Even-Year Asian Pink Salmon (*Oncorhynchus gorbuscha*) Genetic Baseline to Determine Accuracy and Precision of Stock Composition Estimates

**S. Hawkins**

Auke Bay Laboratory, Alaska Fisheries Science Center,  
National Marine Fisheries Service, National Oceanic and Atmospheric Administration,  
11305 Glacier Highway, Juneau, Alaska, 99801-8626, U.S.A.

**N. Varnavskaya**

Kamchatka Research Institute of Fisheries and Oceanography (KamchatNIRO),  
Petropovlovsk-Kamchatsky, 683602, Russia

and

**J. Pohl and R. Wilmot**

Auke Bay Laboratory



Hawkins, S., N. Varnavskaya, J. Pohl, and R. Wilmot. 1998. Simulations of the even-year Asian pink salmon (*Oncorhynchus gorbuscha*) genetic baseline to determine accuracy and precision of stock composition estimates. N. Pac. Anadr. Fish Comm. Bull. No. 1: 213-219

Electrophoretic analysis of Asian even brood-year pink salmon stocks has shown regional heterogeneity (Noll et al. in review). Hypothetical mixed fisheries were created using data from 24 variable loci from Noll et al. in review. The mixture was analyzed to test the accuracy and precision of this baseline data for potential use in mixed fishery analyses. Thirteen stocks were separated into four management regions: Japan, Sakhalin, eastern Kamchatka, and western Kamchatka. Simulations were varied in sample size, number of loci, and percent regional contribution. The simulated mixtures were analyzed using the Conditional Maximum Likelihood Estimate (MLE). The mean estimate, standard deviation, and coefficient of variation were calculated for standardized comparison by both stock and region. Computed MLEs showed that estimates for the Noll et al. baseline improved in accuracy and precision with increased sample size and retention of important loci. When 24 loci and a minimum of 200 samples in a mixture were used, the baseline was approximately 80% accurate in its ability to distinguish regions from a mixture.



### INTRODUCTION

Pink salmon (*Oncorhynchus gorbuscha*) are distributed around the North Pacific Rim from Japan to northern California, extending into the northernmost reaches of Asia and North America. It is the most abundant salmon species in the North Pacific Ocean (Heard 1991) as well as one of the most abundantly harvested. The commercial importance of this species to coastal countries of the North Pacific is verified by a catch rate of approximately 300,000 metric tons annually. Annual odd-year catch rates are typically larger, on the order of 350,000 metric tons, compared to the average even year rate of 250,000 metric tons (Fishery Statistics, 1995). The value of

the fishery often causes conflict when neighboring nations are unable to satisfactorily divide the catch. The complexity of mixed fishery dispute increases where transboundary rivers exist, or migration routes pass through open waters of neighboring countries.

This study examines the feasibility of using the even-year pink salmon genetic baseline (Noll et al. in review) as a tool in assessing stock composition in mixed-stock fisheries. Stock composition estimates from simulated samples of various mixtures can be used to describe the potential utility of the baseline for estimating the composition of real mixtures.

Because gene flow in pink salmon occurs only between generations spawning biennially, significant genetic divergence has occurred between the even-

and odd-year lines (Salmenkova et al. 1981; Beacham et al. 1985, 1988; Gagalchy 1986, Efremov 1991; Makoyedov et al. 1993) and are separate entities for genetic analyses and management purposes. The scope of this paper will discuss only the even brood-year line.

### METHODS

Data were selected from the even-year pink salmon baseline (Noll et al. in review) which included thirteen 1990 collections: two stocks from river systems on Hokkaido Island, Japan, four from southern Sakhalin Island, five from the western Kamchatkan peninsula, and two from the eastern Kamchatkan peninsula, Russia.

Mixtures were simulated from the baseline allele frequencies using the program SIMULATR developed at the Auke Bay Laboratory (Pella et al. 1996). The program generated series of random samples from the baseline data using the bootstrap procedure, and created mixture samples of specified size and expected stock composition. SIMULATR computed the conditional maximum likelihood estimate (MLE) (Pella and Milner 1987) for each set of baseline and mixture samples it generated. To begin the search for an MLE, the 13 baseline stocks were assumed to contribute equally to each mixture. The program searched until at least 95% of the maximum of the likelihood function was attained.

#### Sample Size

SIMULATR created mixture samples from 24 variable loci in the 36 loci baseline whereby each region, (Japan, Sakhalin, West Kamchatka, and East Kamchatka) was comprised of equal portions of the stocks within that region. One thousand sets of baseline and mixture samples were generated from each region, with a mixture sample size of 50. The simulation was repeated varying the mixture sample size from 50 to 1000, for a total of fifteen mixture series for each region. MLEs were computed for each of the 60,000 sets of samples (15 mixtures x 4 regions x 1000 bootstrap samples). The coefficient of variation was calculated and plotted against the sample size for standardized comparison to describe the effect of sample size on the analysis (Marlowe and Busack 1995).

#### Important Loci

The five most distinguishing loci, *GR-1\**, *sAAT-3\**, *PEP-B1\**, *PGDH\**, and *PEP-D2\** were determined by gene diversity analysis (Noll et al. in review). The earlier series of simulations was modified and repeated with these five most distinct

loci omitted and the remaining 19 loci included. The number of sets of baseline and mixture samples generated per experimental combination was reduced to 100 from 1000 in order to limit computation time. Mixture sample size was limited to 200. MLEs, confidence intervals, and coefficients of variation were calculated to compare the 19- and 24-loci analyses. Additional mixtures were created and MLEs computed for a variety of combinations of loci to determine which set of loci produced best accuracy and precision.

#### Precision and Accuracy

Finally, 40 hypothetical mixtures were created in which the percent contribution from each region was varied. MLEs were computed for the series of simulated mixtures using 24 loci, a mixture sample size of 200, and 100 bootstrap samples. An additional four series of mixtures (11 mixtures in each series) were created for which the percent contribution of each region varied from 0% to 100% in increments of 10%. The mean, standard deviation, confidence interval, and coefficient of variation were calculated to describe the accuracy and precision of stock composition estimation from the baseline.

### RESULTS AND DISCUSSION

The simulations and maximum likelihood estimates (MLEs) were used to examine the utility of the Asian pink salmon baseline (Noll et al. in review) for mixed-stock analyses and related applications.

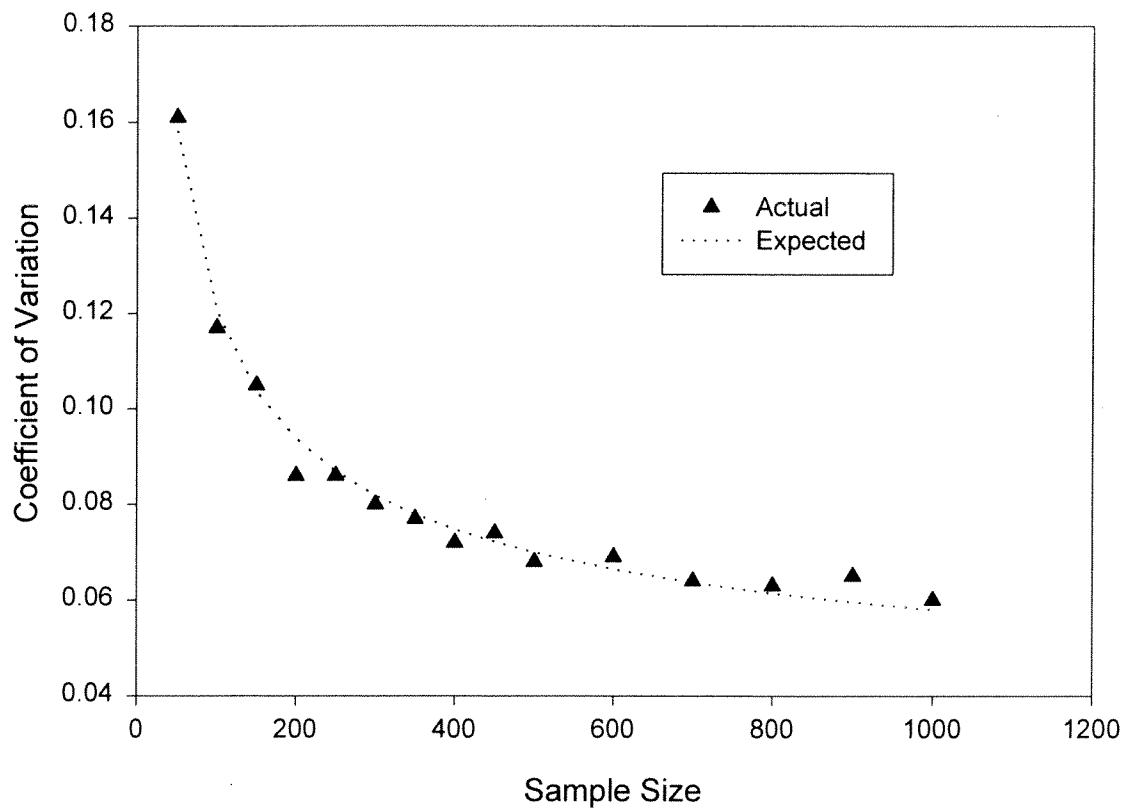
Estimated composition of a mixture improved in accuracy and precision with increased sample size (Table 1), however, significant bias remained in that all stocks were underestimated. The relationship between the coefficient of variation and the mixture sample size showed that a minimum sample size of approximately 200 was necessary for a stock composition estimate with an average standard error of less than 20% (Fig. 1). Rate of change of increased accuracy and precision diminished above  $N=200$ . Estimates obtained from sample sizes less than 200 were of questionable value because the standard error exceeded the actual composition.

MLE's of mixture samples composed of 100% of each of four regions with stocks in equal portions from that region, using 100 bootstrap samples, 24 loci, and a mixture sample size of 200, produced mean estimates (with standard deviations) as follows: Sakhalin - 86% (0.09), Japan - 68% (0.14), West Kamchatka - 84% (0.06), and East Kamchatka - 80% (0.08) (Table 2). True mixtures of multiple stocks will generally yield a more accurate estimate because the MLE was constrained between 0 and 100% and therefore simulated mixtures with contributions

**Table 1. Sample size (N), Estimated Mean, Standard Deviation (S.D.), and Coefficient of Variation (C.V.) where each region comprises 100% of the mixture, using 24 loci.**

N	Japan			Sakhalin			W. Kamchatka			E. Kamchatka		
	Mean	S.D.	C.V.	Mean	S.D.	C.V.	Mean	S.D.	C.V.	Mean	S.D.	C.V.
50	.589	.226	.384	.794	.164	.207	.800	.129	.161	.737	.145	.197
100	.635	.182	.286	.842	.130	.155	.824	.097	.117	.766	.111	.144
150	.667	.153	.230	.859	.100	.117	.833	.087	.105	.778	.098	.126
200	.672	.143	.213	.869	.095	.109	.838	.072	.086	.789	.089	.113
250	.682	.129	.190	.880	.083	.094	.837	.072	.086	.786	.086	.109
300	.687	.119	.174	.886	.073	.083	.841	.067	.080	.790	.082	.104
350	.689	.118	.171	.887	.073	.083	.843	.065	.077	.791	.082	.103
400	.690	.108	.157	.890	.072	.081	.845	.061	.072	.793	.078	.099
450	.692	.110	.159	.895	.067	.075	.842	.063	.074	.790	.079	.100
500	.694	.106	.153	.901	.063	.069	.844	.057	.068	.792	.076	.096
600	.687	.104	.151	.903	.057	.063	.849	.058	.069	.793	.073	.092
700	.695	.100	.143	.906	.060	.066	.850	.054	.064	.794	.070	.088
800	.698	.098	.140	.907	.054	.060	.848	.054	.063	.796	.070	.088
900	.697	.100	.144	.908	.054	.059	.847	.055	.065	.799	.068	.085
1000	.698	.095	.135	.908	.053	.058	.852	.051	.060	.797	.067	.084

**Fig. 1** Curve describing the relationship between sample size and the coefficient of variation. The curve represents a mixture comprising a 100% contribution of the western Kamchatka region and 1000 bootstrap resamplings of the baseline for 24 loci.



**Table 2.** For the 19 loci analysis, sAAT-3\*, GR-1\*, PEP-B1\*, PGDH, and PEP-D2\* were eliminated; for the 18 loci analysis, an additional discriminating locus was removed: FDHG\* from Japan, TPI-4\* from Sakhalin, GPI-A\* from W. Kamchatka, and mIDHP-1\* from E. Kamchatka. Each region consists of 100% of the mixture; the sample size 200.

# Loci	Japan			Sakhalin			W. Kam			E. Kam		
	Mean	S. D.	C.V.	Mean	S. D.	C.V.	Mean	S. D.	C.V.	Mean	S. D.	C.V.
24	.679	.143	.211	.858	.098	.114	.839	.065	.077	.802	.080	.099
19	.532	.192	.361	.800	.141	.176	.659	.161	.244	.657	.159	.242
18	.474	.200	.422	.763	.153	.201	.548	.152	.277	.627	.126	.203

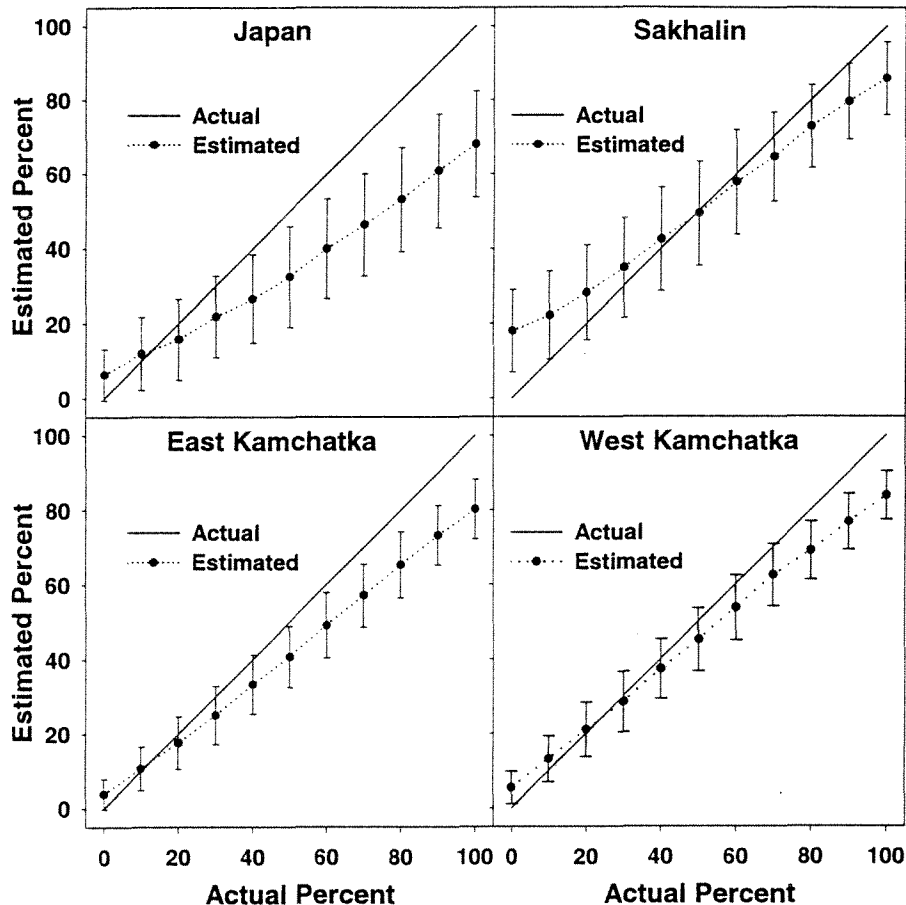
approaching these limits deviate from the actual contribution (Fig. 2). Pella and Milner (1987) noted abundant stocks will be underestimated, and rare stocks, overestimated.

Simulated mixtures of 100% composition of each region produced an overall average maximum likelihood estimate of 80%, with 10% standard deviation, whereas the chum salmon (*O. keta*) baseline (Seeb et al. 1995; Wilmot et al. 1995), and chinook salmon (*O. tshawytscha*) baseline (Marshall et al. 1990; Marlow and Busack 1990) yield estimates of 95% and greater for similar pure (100%) regional

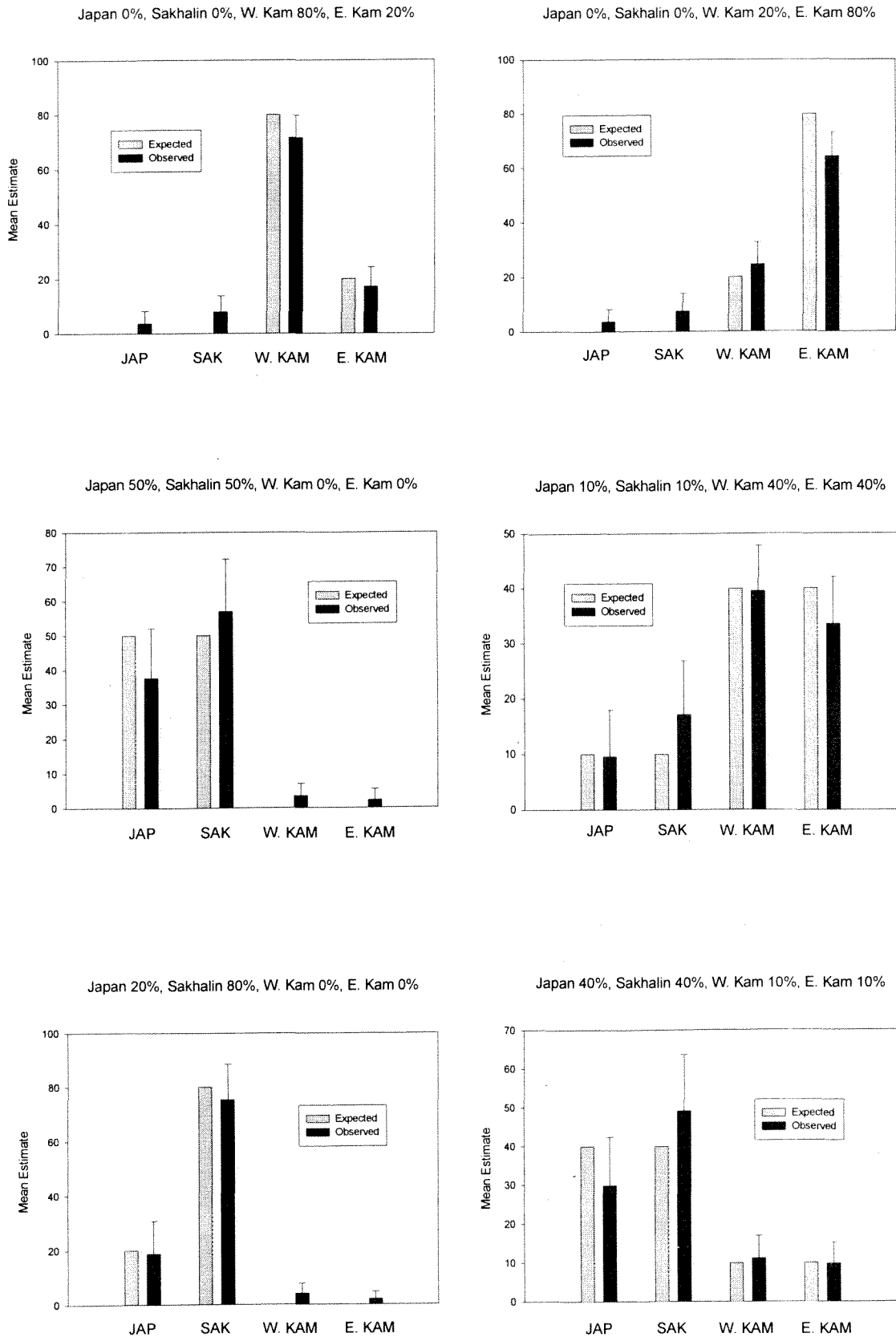
mixtures. Misallocation occurs because stocks are genetically similar.

In mixtures comprised of all four regions, contributions for Sakhalin were overestimated, whereas they were underestimated for Japan, East Kamchatka, and, to a much lesser extent, West Kamchatka. Most of the misallocation went to Sakhalin (Fig. 3), most likely because its populations are the most similar genetically to eastern Kamchatka and Japan (Noll et al. in review). The Japanese region had the largest bias between the actual and the estimated contribution, indicating a lack of

**Fig. 2** Estimated versus true proportions of each region. Estimates are the mean of 100 bootstrap resamplings, and error bars are one standard deviation around the mean.



**Fig. 3** Expected versus observed mean estimates with one standard deviation around the mean for simulated mixtures, using 100 bootstrap resamplings and 24 loci.



differentiating genetic markers. This is confirmed by Noll et al. (In review). The two Japanese populations, Tokushibetsu (northern Hokkaido) and Kushiro (southeastern Hokkaido), are separated by the Kuril Islands and flow into different seas: the Tokushibetsu into the Sea of Okhotsk, and Kushiro into the North Pacific Ocean. They differ genetically (Noll et al., in review.), and need to be considered as separate regions for mixed-stock analysis. While the goal to determine the regional or country of origin of mixtures may be to resolve mixed fishery disputes, Sakhalin Island, Russia and Tokushibetsu, Japan, are genetically one region (Noll et al. in review), making separation difficult with current information.

An increased number of loci might demonstrate greater genetic heterogeneity among regions. The earliest protein studies of Asian pink salmon resolved only 5-6 loci (Zhitovitsky et al. 1989) and indicated that no heterogeneity existed among regions. With data from new loci, heterogeneity among populations has been detected (Noll et al. in review). Gene diversity analysis for 24 loci (Noll et al. in review) show the five loci with the highest level of regional heterogeneity are *AAT-3\**, *GR-1\**, *PEPB-1\**, *PGDH\**, and *PEPD-2\**. When these five loci were eliminated from the analyses, precision and accuracy decreased substantially for every region. For example, the coefficient of variation for western Kamchatka increased from 7.7% with 24 loci, to 24.4% with 19 loci (Table 3). In addition, *FDHG\** was important in determining the accuracy of the estimate for Japan, *TPI-4\** for Sakhalin, *IDHP-1\** for eastern Kamchatka, *GPI-A\** and, to a lesser extent *TPI-4\** and *IDHP-1\**, for western Kamchatka. The omission of six important loci increased the coefficient of variation from 21.1% with 24 loci to 42.2% with 18 loci for Japan (Table 2). Loss of discriminating loci degraded accuracy and precision of composition estimation.

The degradation of proteins in tissue samples of poor quality can result in loss of important loci for composition estimation. Because several loci showing significant regional heterogeneity are expressed in highly labile proteins, sample quality is critical for genetic stock separation. Data for the baseline (Noll et al. in review) included 36 loci out of a potential 69 loci for pink salmon.

### CONCLUSIONS

Each genetic baseline must be evaluated individually for reliability for mixed stock fishery analysis. Conclusions in this manuscript can only be applied to the Noll et al. (in review) baseline.

Precision and accuracy of regional composition improved substantially for mixture sample sizes of about 200. Mixture samples less than 200 were

unsatisfactory because the standard error of the estimates exceeded known contributions.

Maximum likelihood estimates of the Noll et al. (in review) genetic baseline data produced a mean estimate of 80%  $\pm$  10%, with a mixture sample size of 200. Precision may not be accurate enough to estimate regional contributions to mixed stock fisheries. However, it may be the most accurate and cost effective method currently available. Improved sample quality might produce additional distinct loci and improve accuracy and precision of stock contribution estimates.

### REFERENCES

- Beacham T.D. et al. 1985. Biochemical genetic stock identification of pink salmon (*Oncorhynchus gorbuscha*) in southern British Columbia and Puget Sound. Can. J. Fish Aquat. Sci. 42:1474-1483.
- Beacham, T.D. et al. 1988. Variation in body size, morphology, egg size and biochemical genetics of pink salmon in British Columbia. Trans. Amer. Fish Soc. 117:109-126.
- Efremov, V.V. 1991. Homing and population organization of pink salmon. Biologia Moria. 1:3-12. In Russian, English abstract.
- Fishery Statistics, Food and Agriculture Organization yearbook. 1995. FAO Fisheries series No. 48, FAO Statistics series No. 134, p. 165.
- Gagalchy, N. G. 1986. Biochemical polymorphism of Kamchatkan pink salmon, *Oncorhynchus gorbuscha* Walbaum. Genetika. 22:12:2851-2857.
- Heard, W.R. 1991. Life history of pink salmon (*Oncorhynchus gorbuscha*). In: Groot, C., and L. Margolis. Pacific Salmon Life Histories. UBC Press, Vancouver. p. 119-230.
- Makoyedov, A.P. et al. 1993. Population genetic studies of pink salmon, spawning in the rivers of north-eastern Russia. Genetika 29:1366-1374.
- Marlowe, C. and C. Busack. 1995. The effect of decreasing sample size on the precision of GSI stock composition estimates for chinook salmon (*Oncorhynchus tshawytscha*) using data from the Washington Coastal and Strait of Juan de Fuca Troll Fisheries in 1989-90. Northwest Fishery Resource Bulletin, Project Report Series No. 2. Washington Department of Fish and Wildlife, Olympia, WA. 28 pp. (and appendices).
- Marshall, A.R. et al. 1991. Genetic stock identification analysis of three 1990 Washington ocean and Strait of Juan De Fuca chinook salmon fisheries. Washington Department of Fisheries, GSI Summary Report 91-1, Olympia, Washington. 44pp.
- Masuda, M., S. Nelson, and J. Pella. 1991. User's manual for GIRLSEM, GIRLSYM, and

- CONSQRT. Personal computer version. USA-DOC-NOAA-NMFS, Auke Bay Lab., US-Canada Salmon Program, 11305 Glacier Hwy., Juneau, AK 99801-8626. 70 pp.
- Nei, M. 1978. Estimation of heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583-590.
- Noll, C.N. et al. 1997. Genetic relationships among even-year pink salmon (*Oncorhynchus gorbuscha*). In Review.
- Pella, J.J. and G.B. Milner. 1987. Use of genetic marks in stock composition analysis, p. 247-276. In: N. Ryman and F. Utter [ed.] *Population and Fishery Management*. University of Washington Press, Seattle WA.
- Pella, J.J. et al. 1996. Search algorithms for computing stock composition of a mixture from traits of individuals by maximum likelihood. NOAA Tech. Memo NMFS-AFSC-61., 68pp.
- Salmenkova, E.A. et al. 1981. Population genetic differences among next generations of pink salmon, *Oncorhynchus gorbuscha*, spawning in rivers of the Asian coast of the North Pacific. *Genetika and Reproduction of Marine Animals*. Vladivostok. DVNC AN USSR. p. 95-104.
- Seeb, L.W. et al. 1995. Progress report of genetics studies of Pacific Rim chum salmon and preliminary analysis of the 1993 and 1994 South Unimak June fisheries. Regional Information Report 5J95-07, Alaska Department of Fish and Game, Division of Commercial Fisheries Management and Development, Anchorage, AK.
- Wilmot, R.L. et al. 1995. Preliminary results on the origin of chum salmon harvested incidentally in the 1994 Bering Sea Trawl Fishery determined by genetic stock identification. (NPAFC DOC 132) Auke Bay Laboratory, Alaska Fisheries Science Center, NMFS, NOAA, 11305 Glacier Highway, Juneau, AK 99801. 23 pp.
- Zhivotovsky, L.A., M.K. Glubokovsky, R.M. Victorovsky, A.M. Bronevsky, K.I. Afanasev, V.V. Efremov, L.N. Ermolenko, B.A. Kalabushkin, V.G. Kovalev, A.N. Makoedov, T.V. Malinina, S.P. Pustovoit, and G.A. Rubstova. 1989. Genetic differentiation of pink salmon. (English transl.) *Genetika* 25:1261-1274.