Report of the Genetics Workgroup

Pacific Salmon Commission

600-1155 Robson Street
Vancouver, BC, V6E 1B5

October 2007

GENETICS WORKGROUP MEMBERS

W. Stewart Grant (coordinator)
Department of Biological Sciences
University of Alaska Anchorage
Anchorage, AK 99506

Terry Beacham
Pacific Biological Station
Department of Fisheries and Oceans
Nanaimo, BC, V9R 5K6

Craig Busack (Steering Committee)
Washington Department of Fisheries
and Wildlife
600 Capitol Way N.
Olympia, WA 98501-1091

John C. Clark (Steering Committee)
Alaska Department of Fish and Game
P.O. Box 115526
1255 W. 8th Street
Juneau, AK 99811-5526

Sara Gilk
Gene Conservation Laboratory
Alaska Department of Fish and Game
333 Raspberry Road
Anchorage, AK 99518

Chris Habicht
Gene Conservation Laboratory
Alaska Department of Fish and Game
333 Raspberry Road
Anchorage, AK 99518

Steven Kalinowski
Department of Ecology
Montana State University
PO Box 173460
Bozeman, MT 59717-3460

Kathryn Kostow
Ocean Salmon and Columbia River Programs
Oregon Department of Fish and Game
17330 SE Evelyn Street
Clackamas, OR 97015

Steve Latham
Pacific Salmon Commission
600-1155 Robson Street
Vancouver, BC, V6E 1B5

Kristi Miller
Pacific Biological Station
Department of Fisheries and Oceans
Nanaimo, BC, V9R 5K6

Shawn Narum
Columbia River Inter-Tribal Fish
Commission
729 NE Oregon Street, Suite 200
Portland, OR 97206

Christian Smith
Abernathy Fish Technology Center
US Fish and Wildlife Service
1440 Abernathy Creek Road
Longview, WA 98632

Kenneth I. Warheit
Washington Department of Fisheries
and Wildlife
600 Capitol Way N.
Olympia, WA 98501-1091

PREFACE

As a follow-up to the work of the Expert Panel on 'The Future of the Coded Wire Tag Program for Pacific Salmon' in 2005, two workshops[1] were convened to develop recommendations for the integration of GSI information into a coordinated coast-wide management system. The goal was to improve the ability of ocean fisheries to access abundant stocks within impact constraints established for other specific stocks. Workshop participants were directed to identify and quantify costs, implementation steps and the timeframe to implement recommendations.

Within these overall objectives, the specific charge to the Genetics Workgroup (WG) was to develop specific proposals from the following charges: 1) Recommend additional sampling locations, sample sizes, and field and laboratory protocols to improve the GSI database; 2) Recommend how best to incorporate GSI data into ocean salmon management models and regimes; 3) Suggest further research to more effectively incorporate GSI data into the management of ocean fisheries of salmon. The GW focused largely on the first objective because of the lack of time to interact constructively with modelers and managers. Within this more limited focus, the GW spent most of its time in lively discussions of the relative merits of using microsatellites and single nucleotide polymorphisms (SNPs) for GSI. Also included were discussions of ways to enhance the accuracy of GSI estimates with improvements of the statistical treatment of reporting groups and of GSI estimation procedures.

[1] Portland, Oregon 15–17 May 2007 and Vancouver, British Columbia 11–13 September 2007.

GLOSSARY

ADFG	Alaska Department of Fish and Game
CTC	Chinook Technical Committee (PSC)
CTC	Chinook technical committee
CWT	Coded wire tag
DGO	Department of Fisheries and Oceans, Canada
DNA	Deoxyribonucleic acid
ESA	U.S. Endangered Species Act
ESE	Expected squared error
EST	Expressed sequence tags (short DNA sequences)
FOIA	US Freedom of Information Act of 1986
GAPS	Genetic analysis of pacific salmonids
GSI	Genetic stock identification
PCR	Polymerase chain reaction (method of amplifying DNA sequences)
PSC	Pacific Salmon Commission
QTL	Quantitative trait loci
SN	Statistical network
SNP	Single nucleotide polymorphism
TRT	Technical recovery team
WCVI	West Coast of Vancouver Island

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

LIST OF APPENDIX INFORMATION BOXES

PART I. BACKGROUND INFORMATION

BRIEF HISTORY OF GENETIC STOCK IDENTIFICATION

The history of molecular markers in Pacific salmon research and management reaches back to the 1960s, when blood types were used to distinguish populations of sockeye salmon in rivers draining into Bristol Bay (Ridgway and Utter 1964) and to identify major stock components in ocean fisheries (Ridgway 1964; Utter *et al*. 1966). These blood-group polymorphisms, detected with rabbit serum antibodies, were complex, and the salmon red-blood cells could be stored for only a short time (Hodgins 1972). These attempts to use blood-group polymorphisms represented the first use of genetic markers to identify stock components in a mixed-stock fishery.

The appearance of electrophoretic methods to detect protein variants represented a breakthrough in the search for a suitable molecular marker. Allozyme genotypes reflected Mendelian variation and were easy to screen with starch-gel electrophoresis. Allozyme markers were described in Pacific salmon in the late 1960s and early 1970s (Hodgins *et al*. 1969; Utter *et al*. 1970; Utter and Hodgins 1970) and soon after, found their way into a variety of applications, including the reconstruction of the Pacific salmon family tree (Utter et al. 1973a), reproductive and population biology (Utter *et al*. 1973b; Wilmot 1974), and fishery management (Utter *et al*. 1976).

The mid 1970s marked a significant advance in statistical methods to use allozyme variants for mixed-stock analysis. Projects had been initiated on chum salmon populations in Puget Sound (Seeb and Wishart 1977) and on sockeye salmon in Cook Inlet, Alaska (Grant *et al*. 1980) to tease apart components in mixed-stock fisheries. George Milner wrote a maximum likelihood program for mixed stock analysis using the EM algorithm and kindly guided colleagues through the analysis of allozyme data. The early version of the program required as much as 30% of University of Washington's IBM mainframe computing power for some runs.

The 1980s and 1990s witnessed the implementation of GSI capabilities by several laboratories and a growing use of GSI by management (Beacham *et al*. 1985a,b; Shaklee *et al*. 1990; Utter *et al*. 1987). This activity stimulated substantial improvements in GSI estimation procedures (Fournier et al. 1984; Pella and Milner 1987; Wood *et al.* 1987; Pella *et al.* 1996; Pella and Masuda 2001). One important improvement was implementation of Bayesian methods to use previous information to jump-start iterations toward the resolution of stock proportions in a fishery sample. Another advance was to use mixture information to improve the population allele-frequency estimates in the baseline (Pella and Masuda 2001).

Several DNA markers have been considered for GSI over the past 30 years including restriction fragment length polymorphism (RFLP) analysis of mitochondrial DNA (Potter *et al.* 1975; Avise *et al.* 1979; Avise 1989), minisatellites (Jeffreys *et al*. 1985; Galvin *et al.* 1995), random amplified polymorphic DNA (RAPD) (Welsh and McClelland 1990; Williams *et al*. 1990), short interspersed nuclear elements (SINEs) (Okada 1991), amplified fragment length polymorphisms (AFLPs) (Vos *et al*. 1995), and microsatellites (Tautz 1989). The DFO laboratory at Nanaimo, BC spearheaded the use of minisatellites in GSI applications of Pacific salmon, but minisatellites were not cost-effective for GSI and were not widely adopted (Beacham *et al*. 1996a,b; Miller *et*

al. 1996). Microsatellites eventually displaced allozymes as a standard tool for Pacific salmon research when high throughput methods became available (O'Connell and Wright 1997).

Single nucleotide polymorphisms (SNPs) are now gaining popularity in some circles and await evaluation as a population marker. This class of marker represents a shift in focus from protein products or DNA segments to individual nucleotide sites. SNPs ultimately are the sources of variability for most molecular markers. They potentially can provide a wealth of markers, as they occur at frequencies of 1 in 300–500 nucleotide sites throughout the genome (Nielsen 2000). The present generation of SNP assays now provides the capability of screening thousands of SNPs for markers of diseases and quantitative traits. Although SNPs were described some time ago (Botstein *et al*. 1980; Fischer and Lerman 1983), they have been applied only recently to population surveys of Pacific salmon (Smith *et al* 2005).

Three criteria have been used to evaluate new population markers as they have appeared. The first is whether a new marker provides greater population resolution than existing markers. The second is the availability of high throughput assays. This is especially important for GSI applications requiring the analyses of thousands of fish annually. A third criterion is the compatibility of a new marker with the theoretical framework used to make demographic inferences. As SNPs have come into greater use in numerous settings, statistical and theoretical investigations of SNP-based inferences have followed (Kuhner *et al.* 2000; Nielsen 2000; Schlötterer and Harr 2002; Wooding and Rogers 2002; Brumfield *et al*. 2003; among many others).

One potential problem in adopting SNPs arises from the practice of 'high-grading'—choosing only markers that show large differences among populations (for example, humans, Paschou *et al.* 2007). One mechanism producing high-graded markers is natural selection, which can leave a characteristic geographical imprint on populations (*e.g.* Verrelli and Eames 2001). While high-graded SNPs may provide greater resolution among populations for GSI (Smith *et al.* 2005), they may be poorly suited to other applications in fishery management. For example, high-graded SNPs yield biased estimates of genetic diversity for conservation or of population parameters, including effective population size, past demographic events (bottlenecks in population size, founder events), and gene flow ('straying').

INTRODUCTION TO WORKGROUP REPORTS

Members of the genetics workgroup wrote reports on seven key issues between workshops and discussed the results of these reports at the second workshop. These reports appear in the appendix.

Appendix A. Choice of Marker Types for Genetic Stock Identification

One focus of the workgroup was on whether single nucleotide polymorphisms (SNPs) could complement, or possibly replace, microsatellites for GSI. This topic is explored by Smith *et al*., who compare the characteristics of microsatellites and SNPs and attempt to capture the diversity of opinion voiced at the workshops. Preliminary comparisons of regional microsatellite and SNP datasets fail to show clear advantages of one marker type over the other. Cost-benefit analyses of these marker types in coast-wide applications have yet to be conducted and are at the top of the list of recommended actions.

Appendix B. Status of sample collections and genotypic data

One way of increasing population resolution is to boost sampling effort to improve population baselines. Habicht *et al*. summarize the status of sample collections and available population data for the various species of Pacific salmon, but with a focus on coho and sockeye salmon. Presently, microsatellite and SNP data are available for tens of thousands of Chinook and sockeye salmon over a wide geographic range, and large databases are available for chum salmon for both microsatellites and SNPs. This summary will guide future efforts to sample populations in under-represented areas.

Appendix C. Coast-wide integration of GSI data collection, interpretation and use in mixed stock analysis

Coast-wide GSI applications require locus and allele standardization and data sharing among laboratories. Cooperation among laboratories has been excellent for standardizing microsatellite loci and alleles in the GAPS Chinook salmon database and will be essential for on-going standardizations of coho and sockeye salmon databases. An essay by Grant *et al*. reviews the steps needed to establish a standardized database and the development of mechanisms to facilitate the sharing of unpublished data for coast-wide mixed stock analyses. The Logistics Workgroup also explored procedures for data sharing and archiving. The two contributions together provide a detailed picture of how standardization and data sharing can proceed for additional species of Pacific salmon.

Appendix D. Individual assignments and stock composition estimates for mixture analysis

An important way of enhancing the precision of mixed-stock estimates is to refine statistical GSI procedures. Three reports address this issue. The first contribution by Jerry Pella evaluates the use of individual assignments and summing *versus* the estimation of stock proportions (mixture modeling) for GSI estimation. While individual assignments may be valuable for some non-GSI applications, they provide less precision for GSI than does mixture modeling.

Appendix E. Stock aggregation methods

Another approach to improving GSI estimation is to aggregate similar stocks into reporting groups. A report by Warheit *et al*. addresses the question of how best to aggregate stocks. The somewhat subjective methods used by some management and conservation biologists to aggregate populations often place genetically dissimilar populations into one reporting group. A method of statistical networks with an underlying phylogenetic approach to stock aggregation is developed and applied to Chinook salmon populations in Puget Sound as an example. These aggregations produce greater GSI accuracy than did other population groupings used for management and conservation.

Appendix F. Sources of GSI error

Improvement in GSI precision can be greatly enhanced by identifying and reducing particular sources of error in GSI estimates. Steven Kalinowski devised an algorithm to identify major sources of error in GSI estimates in a Chinook salmon fishery off southeastern Alaska. Among three variables investigated (fishery sampling, genotyping, and baseline sampling), uncertainties in baseline allele frequencies represented the largest proportional source of error in this fishery. Uncertainties in baseline allele frequencies can be addressed by greater sampling effort and by statistical procedures accounting for systematic errors in allele-frequency estimation.

Appendix G. Variation in Chinook salmon migration

Regional and temporal GSI estimates collectively can provide insights into within- and between-season variability in run timing and annual shifts in migration patterns. Terry Beacham presents an in-depth examination of in-season monthly Chinook salmon GSI data extending from 2002 to 2006. An important conclusion from these in-season and inter-annual comparisons is that sparse sampling during a fishing season may give a misleading view of the presences of various stocks contributing to a fishery.

CITATIONS

Avise, J.C. 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* 43: 1192–1208.

Avise, J.C., Giblin-Davidson, C., Laerm, J., Patton, J.C., Lansman, R.A. 1979. Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. *Proceedings of the National Academy of Science, USA* 76: 6694–6698.

Beacham, T., Withler, R., Gould, A. 1985a. Biochemical genetic stock identification of chum salmon (*Onchorhynchus keta*) in southern British Columbia. *Canadian Journal of fisheries and Aquatic Sciences* 42: 437–448.

Beacham, T., Withler, R., Gould, A. 1985b. Biochemical genetic stock identification of pink salmon (*Onchorhynchus gorbuscha*) in southern British Columbia. *Canadian Journal of fisheries and Aquatic Sciences* 42: 1474–1483.

Beacham, T., Withler, R.E., Stevens, T.A. 1996a. Stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) using minisatellite DNA variation. *Canadian Journal of Fisheries and Aquatic Sciences* 53: 380–394.

Beacham, T., Miller, K.M., Withler, R.E. 1996b. Minisatellite DNA variation and stock identification of coho salmon. *Journal of Fish Biology* 49: 411–429.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32: 314–31.

Brumfield, R.T., Beerli, P., Nickerson, D.A., Edwards, S.V. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *TREE* 18: 249–256.

Fischer, S.B., Lerman, L.S. 1983. DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: Correspondence with melting theory. *Proceedings of the National Academy of Science, USA* 80: 1579–1583.

Fournier, D.A., Beacham, T.D., Riddell, B.E., Busack, C.A. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. *Canadian Journal of Fisheries and Aquatic Sciences* 41: 400–408.

Galvin, P., McKinnell, S., Taggaart, J.B., Ferguson, A., O'Farrell, M., Cross, T.F. 1995. Genetic stock identification of Atlantic salmon using single locus minisatellite DNA profiles. *Journal of Fish Biology* 47 (Suppl. A): 186–199.

Hodgins, H.O. 1972. Serological and biochemical studies in racial identification of fishes. In *The stock concept in Pacific salmon* (RC Simon, P Larkin, eds), pp. 199–208. University of British Columbia, Vancouver: HR MacMillan Lectures in Fisheries

Hodgins, H.O., Ames, W.E., Utter, F.M. 1969. Variants of lactate dehydrogenase isozymes in the sera of sockeye salmon (*Oncorhynchus nerka*). *Journal of the Fisheries Research Board of Canada* 26: 15–19.

Jeffreys, A.J., Wilson, V., Thein, S.L. 1985. Hypervariable "minisatellite" regions in human DNA. *Nature* 314: 67–73.

Kuhner, M.K., Beerli, P., Yamato, J., Felsenstein, J. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156: 439–447.

Miller, K.M., Withler, R.E., Beacham, T.D. 1996. Stock identification of coho salmon (*Oncorhynchus kisutch*) using minisatellite DNA variation. *Canadian Journal of Fisheries and Aquatic Sciences* 53: 181–195.

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 1534: 931–942.

O'Connell, M., Wright, J.M. 1997. Microsatellite DNA in fishes *Reviews in Fish Biology and Fisheries* 7: 331–363.

Okada, N. 1991. SINEs. *Current Opinion in Genetics and Development* 1: 498–504.

Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W., Drineas, P. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics* 3(9): e160. doi:10.1371/journal.pgen.0030160

Pella, J.J., Milner, G.B. 1987. Use of genetic marks in stock composition analysis. In *Population Genetics and Fishery Management* (N. Ryman, F.M. Utter, eds), pp. 247–276. Seattle: University of Washington Press.

Pella, J., Masuda, M., Nelson, S. 1996. *Search algorithms for computing stock composition of a mixture from traits of individuals by maximum likelihood*. U.S. Department of Commerce, NOAA Tech. Memo, NMFS-AFSC-61, 68p.

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

Potter, S.S., Newbold, J.E., Hutchison III, C.A., Edgell, M.H. 1975. Specific cleavage analysis of mammalian mitochondrial DNA. *Proceedings of the National Academy of Sciences, USA* 72: 4496–4500.

Ridgway, G.J. 1964. Salmon serology. *International North Pacific Fish Commission, Annual Report* 1962: 107–110.

Ridgway, G.J., Utter, F.M. 1963. Salmon serology. *International North Pacific Fish Commission,* Annual Report 1961: 106–108.

Ridgway, G.J., Utter, F.M. 1964. Salmon serology. *International North Pacific Fish Commission,* Annual Report 1963: 149–154.

Schlötterer, C., Harr, B. 2002. Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*. *Molecular Ecology* 11: 947–950.

Shaklee, J.B., Busack, C., Marshall, A., Miller, M., Phelps, S.R. 1990. The electrophoretic analysis of mixed-stock fisheries of Pacific salmon. In *Isozymes: Structure, function, and use in biology and medicine* (Z-I Ogita, CL Markert, eds), pp. 235–265. New York: Wiley Liss.

Smith, C.T., Templin, W.D., Seeb, J.E., Seeb, L.W. 2005b. Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of U.S. and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management* 25: 944–953.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research* 17: 6463–6471.

Utter, F.M., Hodgins, H.O. 1970. Phosphoglucomutase plymorphism in sockeye salmon. *Comparative Biochemistry and Physiology* 36: 195–199.

Utter, F.M., Allendorf, F.W., Hodgins, H.O. 1973a. Genetic variability and relationships in Pacific salmon and related trout based on protein variations. *Systematic Zoology* 22: 257–270.

Utter, F.M., Allendorf, F.W., May, B. 1976. The use of protein variation in the management of salmonid populations. *Transactions of the 41st North American Wildlife and Natural Resources Conference* 41: 373–384.

Utter, F.M., Ames, W.E., Hodgins, H.O. 1970. Transferrin polymorphism in coho salmon (*Onchorhynchus kisutch*). *Journal of the Fisheries Research Board of Canada* 27: 2371–2373.

Utter, F.M., Ridgway, G.J., Ames, W.E. 1966. An examination of the contribution of areas above and below Wood Canyon on the Copper River to the commercial fishery by blood grouping methods. Final report on: Memorandum of Agreement–Assistance in Copper River salmon racial studies to be provided to Region 5 by Region 1, BCF, July and August of fiscal year 1965, 8 pp.

Utter, F.M., Hodgins, H.O., Allendorf, F.W., Johnson, A.G., Mighell, J.L. 1973b. Biochemical variants in Pacific salmon and rainbow trout: their inheritance and application in population studies. *Genetics and mutagensis of fish* (Schroder, J.H., ed.), pp. 329–339. Berlin: Springer-Verlag.

Utter, F.M., Teel, D., Milner, G., McIsaas, D. 1987. Genetic estimates of stock composition of 1983 Chinook salmon , *Oncorhynchus tshawytscha*, harvests off the Washington coast and the Columbia River. *Fishery Bulletin* 85: 13–23.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23: 4407–4414.

Welsh, J., McClelland, M. 1990. Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research* 18: 7213–7218.

Williams, J.G.K., Kubelik, A.R., Livak, K.I., Rafalski, J.A., Tingey, S.V. 1990. DNA polymorphisms amplified by arbitrary primers are useful genetic markers. *Nucleic Acid Research* 18: 6531–6535.

Wilmot, R.L. 1974. A genetic study of the red-band trout (*Salmo* sp.). PhD Thesis, 60 pp., Corvallis: Oregon State University.

Wood, C., McKinnell, S., Mulligan, T., Fournier, D. 1987. Stock identification with the maximum-likelihood mixture model: sensitivity analysis and applications to complex problems. *Canadian Journal of Fisheries and Aquatic Sciences* 44: 866–881.

Wooding ,S., Rogers, A. 2002. The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics* 161: 1641–1650.

PART II. EXECUTIVE SUMMARY

Two workshops brought together experts to evaluate state-of-the-art developments of genetic stock identification (GSI)[2]. The genetics workgroup (one of four) discussed several key issues in three broad topics. The first of these topics dealt with marker development, baseline sampling, standardization and data sharing. The second topic dealt with statistical treatments of genotypic data and included an evaluation of individual assignments *versus* proportion estimation of stock proportions, an examination of how stocks can be aggregated to improve GSI precision, and an assessment of major sources of error in GSI estimation. The last topic was on the use of GSI results to provide insights into Pacific salmon run timing and migration patterns.

MAJOR FINDINGS

Growing use of genetic markers in fishery management

Finding 1. *Genetic markers have been used in numerous facets of fishery management over the past 40 years. The use of genetic markers has increased as new markers provide greater population resolution and ease of screening. The continuing development of statistical methods has provided greater accuracy for GSI estimation and for gaining insights into the population structures and for evaluating conservation status of Pacific salmon populations.*

Choice of marker types for genetic stock identification

Finding 2. *Three criteria are used formally or informally to evaluate the usefulness of new markers as they appear.*

a. *A new marker should provide equal or greater resolution of population differences than existing markers.*

b. *High throughput genotyping should be available to support applications often requiring the analysis of thousands of fish annually.*

c. *A large-scale adoption of a marker by laboratories requires that it be suitability to continue a well-established tradition of research on salmon population biology.*

Finding 3. *Population resolution is influenced by several factors.*

a. *For selectively neutral and unbiased alleles, power depends on the number of independent alleles.*

[2] The term 'genetic stock identification' includes a broad spectrum of genetic applications, but has become synonymous with 'mixed stock analysis'.

b. *The use of 'high-graded markers— markers showing large differences between populations—produces greater than expected resolution from the independent-alleles rule. This greater discriminating power can be due to the use of 'neutral' alleles showing greater than average differences among populations, or to the use of alleles influenced by natural selection.*

Finding 4. *High throughput assays are available for both SNPs and microsatellites to facilitate rapid sample turnaround for in-season management. The automation of genotyping reduces human error. Opportunities for automated genotyping may be greater for SNPs than for microsatellites, but this issue remains unresolved.*

Finding 5. *Presently, genotyping costs per locus appear to be higher for microsatellites than for SNPs, whereas genotyping costs per allele are higher for SNPs. Importantly, genotyping costs to achieve a particular level of population resolution are unknown.*

Finding 6. *High-graded genetic markers may be unsuited to other applications commonly used in fishery management, including the estimation of genetic diversity for conservation or of demographic parameters, such as effective population size, past demographic events (bottlenecks in population size, founder events) and gene flow (straying).*

Finding 7. *Pacific salmon population geneticists are at the forefront of exploring the large-scale use of SNPs in the fishery management. No other group of fishery geneticists can add to the expertise of participants who attended the two workshops.*

Finding 8. *Empirical comparisons of the cost-benefit relationship between microsatellite and SNPs are a high priority and must precede recommendations on marker selection.*

Status of markers and samples of coho and sockeye salmon

Chinook salmon

Finding 9. *The development of a coast-wide GAPS microsatellite baseline for Chinook salmon represented a considerable advance over the use of allozyme baselines by making data readily accessible over the internet. Lessons learned from GAPS can be used to develop baselines for other species of Pacific salmon. About 51 SNP assays are available for Chinook salmon. About 25,000 fish have been examined for SNPs in samples from Russia to California. Several thousand Chinook salmon from Southeast Alaska and the Yukon-Kuskokwim rivers have been examined to support transboundary management.*

Coho salmon

Finding 10. *Numerous populations of coho salmon have been surveyed for variability at numerous microsatellite loci (and two MHC loci in some areas) by various agencies. Samples extend from Southeast Alaska to northern California. At least 42 SNP assays have been developed for coho salmon, but only about 400 fish have been examined for variability in samples extending from Russia to Washington.*

Sockeye salmon

Finding 11. *Several regional databases of microsatellite markers have been developed for sockeye salmon. Most surveys have included populations in British Columbia, and a few populations of conservation concern in the Columbia-Snake river drainage. About 35,000 sockeye salmon have been examined for SNP variability in samples extending from Russia to Washington-Idaho, but with a concentration in Alaska around Bristol Bay and the Alaska Peninsula, where this species is most abundant.*

Chum salmon

Finding 12. *Numerous microsatellites have been developed for chum salmon, and numerous populations have been surveyed. About 77 SNP assays have been developed for chum salmon. About 12,000 chum salmon have been genotyped in samples extending from Korea to Washington.*

Pink salmon

Finding 13. *No SNP assays have been developed for pink salmon.*

Coast-wide integration of GSI data collection and use

Finding 14. *The value of GSI is greatly enhanced by ensuring that regional datasets can be merged into a larger coast-wide dataset. Merging data from several laboratories requires attention to four layers of detail.*

a. A common set of loci must be used among laboratories for each class of molecular marker.

b. Laboratories must standardize allelic identities and allelic nomenclature.

Standardization is complicated for microsatellites, because different automated platforms generally produce different allelic mobilities. Rapid standardization of alleles may be achieved with allelic ladders, without the need for exchanging tissues or DNA. Minimal allelic standardization is required for single nucleotide

polymorphisms (SNPs), as only four easily identified nucleotide states are possible at a nucleotide position.

c. Spatial scales of sampling effort must be consistent among laboratories, so that the most important spawning populations contributing to a fishery are sampled.

While allelic identification among laboratories may not be problematic for SNPs, polymorphisms identified in one region may not provide adequate population resolution in another region.

d. Statistical procedures should be consistent among laboratories. The usefulness of coast-wide analyses depends on standardizing these procedures among laboratories.

Finding 15. Sharing of baseline data among laboratories is essential to address coast-wide GSI problems. Data sharing can be hindered by several factors.

a. Protection by researchers of proprietary information for scientific publication.

b. Hesitation among agencies to share data for fear that some interpretations of a dataset may not prove beneficial to their interests.

Different interpretations of the same data can potentially arise from the use of different statistics, or the inclusion of some samples but not others for mixed stock analysis. Data sharing has traditionally depended on the goodwill and cooperation of personnel in these agencies. However, when problems arise among laboratories, cooperation may have to be implemented by memoranda of agreements that clearly outline how shared data can be used.

Individual assignments and stock composition estimates for mixtures

Finding 16. Artificial and natural marks have been useful in salmon management to identify populations of origin of migrating salmon. An important advantage of natural marks is their complete coverage of all stocks and all individuals in the stocks. However, natural marks provide less certainty in the source identification of individuals than do artificial marks.

Finding 17. The relative frequencies of the natural marks differ among populations and provide some information to probabilistically separate mixture individuals to their sources. Both the sources of individuals and the stock composition of the mixture must be estimated and two general approaches to this dual estimation problem are possible.

a. Classical individual assignments *methods have been less commonly applied in fisheries research.*

b. Mixture modeling *methods are more widely used in fisheries research with both frequentist and Bayesian approaches. Mixture modeling is generally superior to the classical individual assignments method for the dual estimation problem. Although the cost of baseline development and of processing sampled mixture individuals for natural marks may be significant, the cost of statistical estimation is negligible.*

Finding 18. *The Bayesian approach extends mixture modeling to include estimation of both the stock composition and the allele relative frequencies in contributing stocks.*

Aggregating Stocks for Improved Genetic Stock Identification

Finding 19. *A comprehensive knowledge of all stocks is unnecessary, if stocks can be aggregated into groups by assuming that stocks in a group share common characteristics that subject the stocks to the same or similar exploitation rates. Similar biology and recency of common ancestry, measured by genetic similarity, should govern how stocks are aggregated.*

Finding 20. *Aggregation schemes inconsistent with genetic relationships among stocks reduce the accuracy and precision of GSI, thereby limit the usefulness of genetic analyses, and compromise the ability to manage fisheries with a full suite of data. The use of phylogenetic methods to identify genetically similar populations increases GSI accuracy. A Statistical Networks procedure (SN) was superior to two other aggregating procedures for identifying monophyletic groups of Chinook salmon populations in Puget Sound.*

Finding 21. *Standard quantitative stock aggregations should be designed coast-wide to be consistent with the phylogenetic relationships of stocks, and to maximize value to address specific fishery management needs.*

Sources of error affecting GSI accuracy

Finding 22. *Several sources of error influence the accuracy of GSI estimation.*

a. Sampling of the fishery. Error arises from small sample sizes and from non-random sampling.

b. Random sampling fails to include all stocks present in the fishery.

c. Sampling a finite number of genetic markers.

d. Genotyping error.

e. Errors on allele frequencies from sampling a finite number of individuals in baseline populations.

f. The inclusion of fish in the fishery sample from populations not in the population baseline also introduces error.

Finding 23. *Partitioning of total expected square error (ESE)—a variance-like variable, including the effect of bias—into components b, c, and e showed that the largest source of error was due to uncertainties in allele frequencies in baseline populations (87.5%). A smaller proportion was due to fishery sampling (9.5%) and a very small proportion is due to genotypic sampling (2.7%). As the fishery used for this study was typical of other fisheries, these results likely show general trends for GSI estimates for other fisheries.*

Finding 24. *When the level of differentiation among baseline populations is low ($F_{ST} < 0.01$), increased sampling will improve GSI accuracy. In other cases, statistical approaches can be used to improve baseline allele frequencies with two approaches.*

a. Generalized expectation maximum algorithm uses mixture samples to improve allele-frequencies estimates in the baseline populations. This approach is incorporated into available Bayesian GSI programs.

b. Spatial models can be used to improve allele-frequency estimates by assuming that nearby populations tend to be similar.

Finding 25. *The magnitude of error from unsampled source populations potentially can be estimated with three approaches.*

a. Simulations to examine the impact of excluding some existing populations from baselines.

b. *Spatially explicit models of population structure can be constructed to estimate allele frequencies of unsampled populations, and these estimates could be used in conventional GSI simulations.*

c. *The Bayesian missing-data model may successfully identify unsampled populations contributing to a fishery.*

Finding 26. *Error in GSI estimation of low-contributing stocks in a mixture is difficult to evaluate, but can be reduced in part by larger fishery sample sizes. Current GSI algorithms tend to bias stock composition estimates toward 1/k, where k is the number of stocks contributing to the baseline.*

Intra-annual and inter-annual variation in stock composition

Finding 27. *GSI estimates provide an opportunity to understand important features of salmon migration and spawning biology. In-season comparisons of GSI estimates in an area off northern British Columbia revealed contrasting abundance trends for fish from different areas. These estimates show shifts in stock compositions of fish in the troll fishery during the fishing season, which likely reflect migration patterns of various stock groups past the Queen Charlotte Islands.*

Finding 28. *Annual comparisons of five major stocks in either the troll or commercial catches showed shifts for some regions but not for others. Annual variation was most pronounced for Oregon fish with the highest proportions in August 2004. Fraser River fish also increased in 2002 and 2006, likely reflecting the strong returns to the Thompson River drainage in those years.*

Finding 29. *Seasonal and annual comparisons are possible only after a large-scale population baseline has been established to identify stocks potentially contributing to a fishery. Seasonal and annual comparisons require frequent sampling during a fishing season to provide an accurate view of changes in contributing stocks.*

MAJOR RECOMMENDATIONS

Maintenance of existing databases

Recommendation 1. *Maintain and improve existing standardized microsatellite population baselines*

a. Existing microsatellite baselines provide the only means of addressing some management problems.

b. These baselines should be maintained and extended to provide greater levels of population resolution.

Recommendation 2. *Support continued development of genetic markers (particularly for SNPs in sockeye salmon coast-wide)*

a. Use appropriate lessons from the GAPS approach to marker standardization for the development of population baselines for additional species.

b. Develop appropriate markers for use in a coast-wide baseline.

Recommendation 3. *Empirical comparisons of SNPs & microsatellites on a coast-wide scale, with focus on Chinook and sockeye*

a. Even though SNPs often provide a high level of resolution for discriminating among regional populations, can they be effective in a coast-wide baseline?

b. The particular sample of SNP or microsatellite loci for a regional comparison can determine the outcome of a comparison. Hence, appropriate marker should be used in a comparison.

c. Simulations can be used to assess the level of resolution that a marker provides to discriminate among a group of populations.

d. Blind samples of known origins should be used in a GSI analysis to examine resolution of marker types.

e. Evaluations between marker types should be posed in terms of cost for a given amount of population resolution, not just the cost of genotyping.

Recommendation 4. *The potential of a marker type to resolve features of salmon population dynamics, in addition to GSI (mixed stock analysis), should be considered before adopting one marker and abandoning another.*

a. Most models of population structure assume the selective neutrality of alleles.

b. High-graded markers showing strong differences among populations may improve GSI estimation, but produce biased estimates of demographic parameters, such as effective population size and gene flow.

Improvement of statistical GSI methods

Recommendation 5. Support studies investigating sources of GSI error. Preliminary results of theoretical and simulation studies point to ways in improving GSI accuracy.

a. Investigate ways of improving allele-frequency estimates of populations in baseline. Only marginal gains in accuracy can be achieved with larger samples of fishery mixtures and genetic markers.

b. Support studies of other sources of GSI error, including upward bias of low-frequency stocks in mixture, and missing baseline populations.

c. Adopt mixture modeling *for GSI estimation.*

Recommendation 6. Re-examine methods used to aggregate baseline stocks into reporting groups to increase GSI accuracy.

Use of GSI to understand ocean migration and abundance patterns

Recommendation 7. Support summary studies of seasonal and multi-year GSI results to better understand the ocean biology of Pacific salmon.

Incorporation of GSI into Pacific salmon population models and harvest management

Recommendation 8. Support collaborations between geneticists and population modelers and harvest managers to enhance the utility of GSI results.

PART III. JUSTIFICATIONS AND RATIONALE FOR EXPERT PANEL FINDINGS AND RECOMMENDATIONS.

MAJOR FINDINGS

Wide-spread use of genetic markers in fishery management

Finding 1. *Genetic markers have been used in numerous facets of fishery management over the past 40 years. The use of genetic markers has increased as new markers provide greater population resolution and ease of screening. The continuing development of statistical methods has provided greater accuracy for GSI estimation and for gaining insights into the population structures and for evaluating conservation status of Pacific salmon populations.*

The history of molecular markers in Pacific salmon research and management reaches back to the 1960s, when blood types were used to distinguish populations of sockeye salmon in Bristol Bay rivers. Since then numerous molecular markers have found their way into a variety of management applications, as well as in systematics, reproductive and population biology, and conservation.

Choice of marker types for genetic stock identification

Finding 2. *Three criteria are used formally or informally to evaluate the usefulness of new markers as they appear.*

A new marker should provide equal or greater resolution of population differences than existing markers.
High throughput genotyping should be available to support applications often requiring the analysis of thousands of fish annually.
A large-scale adoption of a marker by laboratories requires that it be suitability to continue a well-established tradition of research on salmon population biology.

Several DNA markers have been considered for GSI over the past 30 years including restriction fragment length polymorphism (RFLP) analysis of mitochondrial DNA, minisatellites, random amplified polymorphic DNA (RAPD), short interspersed nuclear elements (SINEs), amplified fragment length polymorphisms (AFLPs), and microsatellites. Microsatellites eventually displaced allozymes as a marker of choice for Pacific salmon research and management, when high throughput methods became available.

The focus now is on the relative merits of single nucleotide polymorphisms (SNPs) and microsatellites (Figure 1). SNPs were initially developed to map genetic diseases in the human genome, but are now used for individual identification, pedigree analysis and cultivar selection

in fish breeding programs. Recently, SNPs have been used as a population marker, and hence the present evaluation to assess the suitability of wider use with Pacific salmon. A new population marker should possess three characteristics: 1) equal or greater resolution of population differences than existing markers, 2) high throughput genotyping for applications requiring the analysis of thousands of fish annually, and 3) suitability to continue a well-established tradition of research on salmon population biology. A cost-benefit analysis of these factors is needed before a new marker can be adopted for GSI estimation.

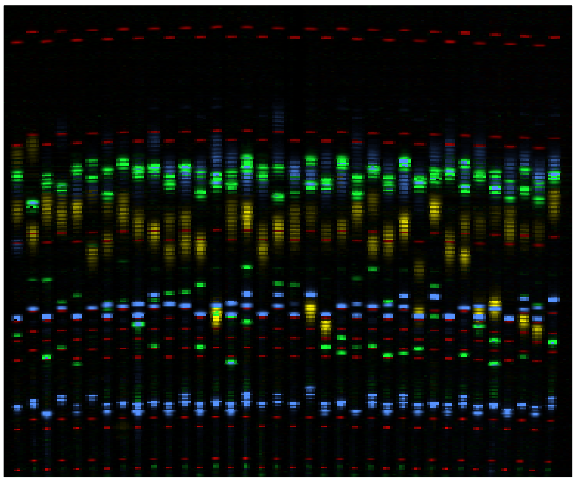

Figure 1. Multiplexed panel of microsatellites.

Finding 3. Population resolution is influenced by several factors.

a. *For selectively neutral and unbiased alleles, power depends on the number of independent alleles.*
b. *The use of 'high-graded markers— markers showing large differences between populations—produces greater than expected resolution from the independent-alleles rule. This greater discriminating power can be due to the use of 'neutral' alleles showing greater than average differences among populations, or to the use of alleles influenced by natural selection.*

Population resolution is influenced by several factors. The statistical power provided by a marker to resolve population differences, or to estimate contributing stocks in a fishery, largely depends on the number of independent alleles, if the markers are not under selection. However, empirical comparisons of SNPs demonstrate that SNPs provide more discriminating power than expected with the independent-alleles rule. The greater discrimination is apparently due to the use of only SNP markers that show large population differences (ascertainment bias). This bias arises from neutral markers showing greater than average differences between populations, or from markers influenced by directional selection. Presently, SNP markers have been used to address regional problems, so the level of resolution among coast-wide populations remains unknown.

Finding 4. *High throughput assays are available for both SNPs and microsatellites to facilitate quick sample turnaround for in-season management. The automation of genotyping reduces human error. Opportunities for automated genotyping may be greater for SNPs than for microsatellites, but this issue remains unresolved.*

Finding 5. *Presently, genotyping costs per locus appear to be higher for microsatellites than for SNPs, whereas genotyping costs per allele are higher for SNPs. Importantly, genotyping costs to achieve a particular level of population resolution are unknown.*

High throughput assays are available for both SNPs and microsatellites to facilitate quick sample turnaround for in-season management. Automation of laboratory procedures and genotype interpretations may improve throughput and reduce genotyping error in both marker types. Although opportunities for automated genotyping may be greater for SNPs (Figure 2) than for microsatellites, this issue remains unresolved. Presently, genotyping costs per locus appear to be higher for microsatellites than for SNPs, whereas genotyping costs per allele are higher for SNPs. Importantly, genotyping costs to achieve a particular level of population resolution are unknown.

Finding 6. *High-graded genetic markers may be unsuited to other applications commonly used in fishery management, including the estimation of genetic diversity for conservation or of demographic parameters, such as effective population size, past demographic events (bottlenecks in population size, founder events) and gene flow (straying).*

Lastly, a new marker should be compatible with the large body of theory used to interpret genotypic data. One hesitation in adopting SNPs arises from the practice of 'high-grading'—choosing only markers showing large differences between populations. One mechanism leading to high-graded markers is natural selection, which can produce a characteristic geographical imprint. While high-graded SNPs may provide greater resolution for GSI, they may be unsuited to other applications commonly used in fishery management. High-graded SNPs cannot provide unbiased estimates of genetic diversity for conservation or of demographic parameters, such as effective population size, past demographic events (bottlenecks in population size, founder events) and gene flow ('straying').

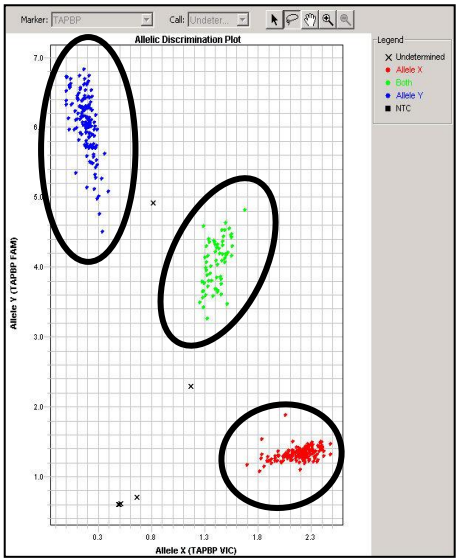

Figure 2. Single nucleotide polymorphism (SNP) panel of genotypes.

Finding 7. *Pacific salmon population geneticists are at the forefront of exploring the large-scale use of SNPs in the fishery management. No other group of fishery geneticists can add to the expertise of participants at the two workshops.*

Pacific salmon population geneticists have historically pioneered the use of genetic markers in fishery management are first in exploring the large-scale use of SNPs in the fishery management. No other group of fishery geneticists can add to the expertise of participants who attended the two workshops. The evaluation of SNPs and microsatellites will have important repercussions, not only for the use of molecular markers in the management of Pacific salmon, but also for applications of molecular markers to major fisheries in other regions of the globe.

Finding 8. *Empirical comparisons of the cost-benefit relationship between microsatellite and SNPs are a high priority and must precede recommendations on marker selection.*

Empirical comparisons of the cost-benefit relationship between microsatellite and SNPs are a high priority and must precede recommendations on marker selection. Cost-benefit relationships, however, change with technological advances and depend on species, laboratory infrastructure, and geographic scale. These analyses require immediate attention.

Status of markers and samples of Chinook, coho, sockeye, and other Pacific salmon

Finding 9. *The development of a coast-wide GAPS microsatellite baseline for Chinook salmon represented a considerable advance over the use of allozyme baselines by making data readily accessible over the internet. Lessons learned from GAPS can be used in the development of baselines for other species of Pacific salmon. About 51 SNP assays are available for Chinook salmon. About 25,000 fish have been examined for SNPs in samples from Russia to California. Several thousand Chinook salmon from Southeast Alaska and the Yukon-Kuskokwim rivers have been examined to support transboundary management.*

Finding 10. *Numerous populations of coho salmon have been surveyed for variability at numerous microsatellite loci (and two MHC loci in some areas) by various agencies. Samples extend from Southeast Alaska to northern California. At least 42 SNP assays have been developed for coho salmon, but only about 400 fish have been examined for variability in samples extending from Russia to Washington.*

Finding 11. *Several regional databases of microsatellite markers have been developed for sockeye salmon. Most surveys have included populations in British Columbia, and a few populations of conservation concern in the Columbia-Snake river drainage. About 35,000 sockeye salmon have been examined for SNP variability in samples extending from Russia to Washington-Idaho, but with a concentration in Alaska around Bristol Bay and the Alaska Peninsula, where this species is most abundant.*

Finding 12. *Numerous microsatellites have been developed for chum salmon and numerous populations have been surveyed. About 77 SNP assays have been developed for chum salmon. About 12,000 chum salmon have been genotyped in samples extending from Korea to Washington.*

Finding 13. *No SNP assays have been developed for pink salmon.*

One step in constructing databases for species of Pacific salmon is the development of regional baselines, usually by agencies with regional management mandates. Coast-wide baselines often follow the development of regional baselines after standardizations of loci and alleles. Previously, coast-wide allozyme baselines were developed and maintained through collaborations and workshops to standardize allelic designations. The development of a coast-wide GAPS microsatellite baseline for Chinook salmon (Figure 3) was a continuation of this process, but represented a considerable advance by making data readily accessible over the internet. As a start toward the development of additional coast-wide baselines, this section summarizes regional and coast-wide datasets with a focus on coho and sockeye salmon. These summaries are snapshots of a growing set of regional and coast-wide databases for SNP and microsatellite markers.

Microsatellite markers

Coho salmon: Numerous populations of coho salmon have been sampled by DFO, Nanaimo in several regions and examined for variability in 13 microsatellite loci and two MHC exons. These samples are concentrated in British Columbia with representative samples from Southeast Alaska and Washington State. A large number of samples have been examined for variability at 11 microsatellite loci by NOAA Fisheries, Seattle, in samples from populations extending from southern British Columbia to northern California. Presently, 61 microsatellite primers developed for coho or other species of salmon have been used to screen for variability in coho salmon.

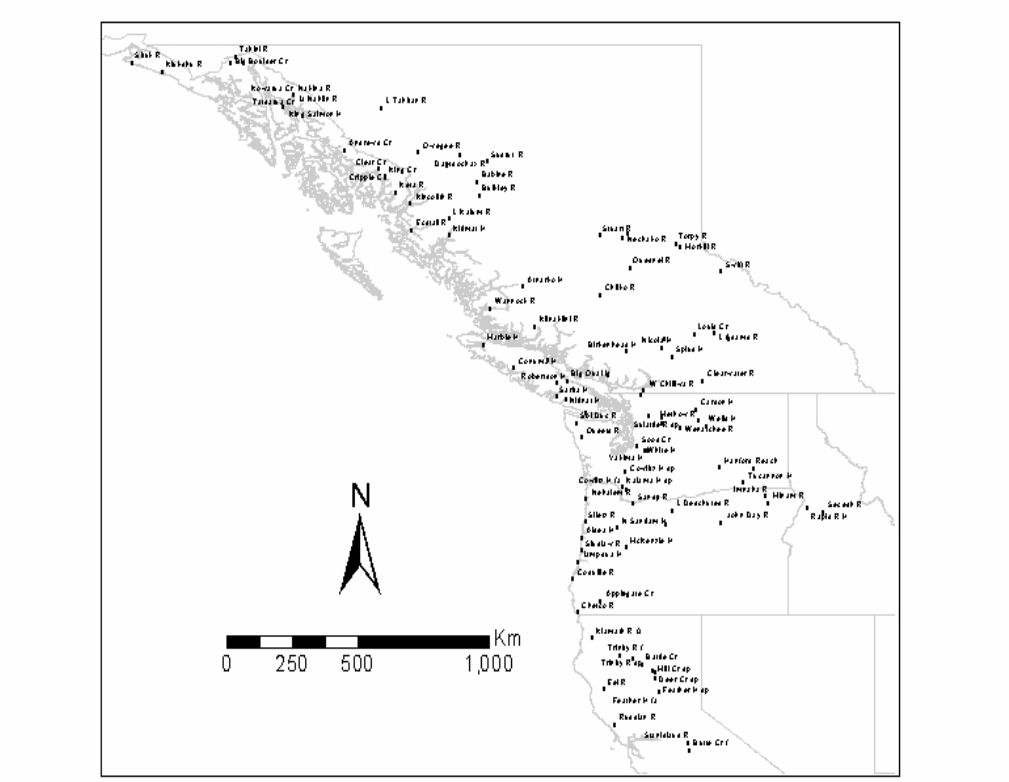

Figure 3. Map showing locations of samples in GAPS baseline.

Sockeye salmon: Several regional databases for microsatellite markers exist for sockeye salmon that have been used by DFO and NOAA. Most surveys of microsatellite loci have been of populations in British Columbia, and of a few populations of conservation concern in the Columbia-Snake river drainage.

Single nucleotide polymorphisms (SNPs)

The number of SNP assays and the number of samples examined are growing rapidly. Most SNP databases encompass only regional sets of populations. Presently, 51 genotypic assays are available for Chinook salmon, 19 for coho salmon, 77 for chum salmon, 44 for sockeye salmon, and none for pink salmon. At least 42 SNP assays have been developed for coho salmon, but only about 400 fish have been examined for variability in samples extending from Russia to Washington.

Regional surveys for some species are quite large. About 35,000 sockeye salmon have been examined for SNP variability in samples extending from Russia to Washington-Idaho, but with a concentration in Alaska around Bristol Bay and the Alaska Peninsula, where this species is most abundant. About 12,000 chum salmon have been genotyped in samples extending from Korea to Washington, and nearly 25,000 Chinook salmon have been examined in samples from Russia to California. Several thousand Chinook salmon from Southeast Alaska and the Yukon-Kuskokwim rivers have been examined to support transboundary management.

Coast-wide integration of GSI data collection, interpretation and use in mixed stock analysis

Finding 14. *The value of GSI is greatly enhanced by ensuring that regional datasets can be merged into a larger coast-wide dataset. Merging data from several laboratories requires attention to four layers of detail.*

- *a. A common set of loci must be used among laboratories for each class of molecular marker.*
- *b. Laboratories must standardize allelic identities and allelic nomenclature.*
- *c. Spatial scales of sampling effort must be consistent among laboratories, so that the most important spawning populations contributing to a fishery are sampled.*
- *d. Statistical procedures should be consistent among laboratories. The usefulness of coast-wide analyses depends on standardizing these procedures among laboratories.*

The value of GSI is greatly enhanced by ensuring that regional datasets can be merged into a larger coast-wide dataset. After the creation of a coast-wide baseline, data should be accessible in a timely manner to management agencies responsible for maintaining sustainable harvests of salmon. Previous efforts to integrate databases for Chinook salmon (GAPS) have proved successful, and this database has provided information to management that would not have been possible with the separate analyses of individual datasets.

Merging data from several laboratories requires attention to four layers of detail. First, a common set of loci must be used among laboratories for each class of molecular marker. Second, laboratories must standardize allelic identities and allelic nomenclature. This is complicated for microsatellites, because different automated platforms generally produce different allelic

mobilities, even in the same laboratory. These two issues can be resolved by collaborations among laboratories and periodic workshops. Alternatively, rapid standardization of alleles may be achieved with allelic ladders, without the need for exchanging tissues or DNA. Minimal allelic standardization is required for single nucleotide polymorphisms (SNPs), as only four easily identified nucleotide states are possible at a nucleotide position.

Third, spatial scales of sampling effort must be consistent among laboratories, so that the most important spawning populations contributing to a fishery are sampled. While allelic identification among laboratories may not be problematic for SNPs, polymorphisms identified in one region may not be present in another region. For example, SNP polymorphisms developed for Alaskan populations may be useful for differentiating Asian populations from North American populations, but may be less useful for distinguishing among Asian populations.

Fourth, statistical procedures should be consistent among laboratories. Sampling design and statistical power influence inferences about population structure, and hence influence the accuracy and utility of GSI. Sampling design is often complicated by the need to resolve run or spawning time components of a population. One goal is to achieve consistency in methods used to aggregate reporting groups for GSI. In addition to the completeness of a population data baseline, the results of mixed-stock analyses depend on the timing and sizes of samples from ocean or river mouth harvests, and on the particular statistical method used to estimate the composition of the mixture (e.g. individual assignment or proportion estimation). The usefulness of coast-wide analyses depends on standardizing these procedures among laboratories.

These four considerations set the stage for the sharing of genetic data to support GSI of fishery samples. The availability of up-to-date, but often unpublished, data is vital to these efforts. Requests for information may include tissue samples for additional analyses, genotypic or allele frequency data, summary statistics or draft reports. While funding agencies may impose data-sharing requirements on researchers, laboratories generally receive support from several in-house and agency sources, each of which may have different data-sharing mandates.

Finding 15. *Sharing of baseline data among laboratories is essential to address coast-wide GSI problems. Data sharing can be hindered by several factors.*

Protection by researchers of proprietary information for scientific publication. Hesitation among to share data for fear that some interpretations of a dataset may not prove beneficial to their interests.

The first step toward facilitating the easy distribution of data is to establish a web-based electronic ‘meta-database’ accessible to stakeholders and management. The primary function of this database would be to catalogue existing primary genetic data (markers, sample dates and sampling localities), biological information (population profiles) and biological materials (tissues, otoliths and scales) available for genetic analysis. A meta-database would also help to improve the designs of research projects and sampling. This database might include the following:

- Existing allozyme, mtDNA, microsatellite, SNP, and EST datasets and where they are located;
- Existing collections of historical biological material that could be used to extract DNA. Archived scales and otoliths can be used to estimate allele frequencies in past populations;
- List of past and current genetics projects, including laboratory location, researchers' names and the natures of the projects;
- Profiles and contact information of active researchers working on the genetics of salmon.

Data sharing can be complicated by other factors. One is the protection by researchers of proprietary information for scientific publication. Another is agencies' hesitation to share data for fear that some interpretations of a dataset may not prove beneficial to their interests. Different interpretations of the same data can potentially arise from the use of different statistics or the inclusion of some samples but not others for mixed stock analysis. Data sharing has traditionally depended on the goodwill and cooperation of personnel in these agencies. However, when problems arise among laboratories, cooperation may have to be implemented by memoranda of agreements that clearly outline how shared data can be used.

Individual assignments and stock composition estimates for a mixture when source marks are not definitive

Finding 16. *Artificial and natural marks have been useful in salmon management to identify populations of origin of migrating salmon. An important advantage of natural marks is their complete coverage of all stocks and all individuals in the stocks. However, natural marks provide less certainty in the source identification of individuals than do artificial marks.*

Artificial and natural marks have been useful in salmon management to identify populations of origin of migrating salmon. An important advantage of artificial marks is that the sources of marked individuals are known with certainty. Therefore, if marking were complete, for instance, the stock proportions from the marks in a fishery sample would be directly observable and these would be the maximum likelihood estimator of the catch stock composition. A disadvantage of artificial marks is their expense in application and in determination of the source at recovery. As a consequence, artificial marking is often incomplete, as neither all stocks nor all individuals in a stock are marked. The lack of marks for some stocks in a mixture is highly problematic for assessing mixed-stock composition.

Natural marks include scale features, parasites, and genotypes. The important advantage of natural marks is that they provide complete coverage of all stocks as well as of all individuals in the stocks. However, natural marks provide less certainty in the source identification of individuals than do artificial marks. The relative frequencies of the natural marks differ among populations and provide some information to probabilistically separate mixture individuals to their sources.

Finding 17. *The relative frequencies of the natural marks differ among populations and provide some information to probabilistically separate mixture individuals to their sources. Both the sources of individuals and the stock composition of the*

mixture must be estimated and two general approaches to this dual estimation problem are possible.

a. Classical individual assignments *methods have been less commonly applied in fisheries research.*

b. Mixture modeling *methods are more widely used in fisheries research with both frequentist and Bayesian approaches. Mixture modeling is generally superior to the classical individual assignments method for the dual estimation problem. Although the cost of baseline development and of processing sampled mixture individuals for natural marks may be significant, the cost of statistical estimation is negligible.*

Finding 18. The Bayesian approach extends mixture modeling to include estimation of both the stock composition and the allele relative frequencies in contributing stocks.

Because the source identity of an individual is almost never certain from its natural marks, both the sources of individuals and the stock composition of the mixture must be estimated. Two general approaches to this dual estimation problem are possible. The first is termed the classical *individual assignments* method, and although not recommended for the dual estimation problem, it has been commonly applied in fisheries research. The classical individual assignments method is used here to motivate and explain the second, and recommended, approach based on *mixture modeling*. Mixture modeling methods are well developed and more widely used in fisheries research with both frequentist and Bayesian versions available. Mixture modeling is generally superior to the classical individual assignments method for the dual estimation problem. Although the cost of baseline development and of processing sampled mixture individuals for natural marks may be significant, the cost of statistical estimation is negligible.

An ostensibly reasonable approach to the dual estimation is the classical individual assignments method. This method includes two steps applied once to the mixture sample: 1) assignment of the mixture individuals to source populations, and 2) estimation of the mixture composition from the assignments using multinomial sampling theory. In the first step, the multilocus genotype for an individual is matched to the population with the most frequent occurrence of the genotype (maximum frequency or MAF rule). Promotion of the MAF rule is misleading because a superior rule, the maximum *a posteriori* or MAP rule, has a lower expected frequency of assignment errors for arbitrary mixtures. The MAP rule assigns each individual to the source stock estimated to contribute the greatest proportion of its genotype to the mixture. The MAF and MAP rules agree only when stocks are equally probable in the mixture. A justification for using the MAF rule may be a lack of information about the mixture composition, but after assignments are completed, some knowledge about stock composition becomes available. In general, the estimated composition from the assignments differs from the assumed equal composition. Additional cycles of assignments and estimation of the mixture composition take advantage of the new knowledge about the mixture composition, but the classical individual assignments method fails to do so. Also, at the second step, the classical individual assignments method, ignores the probable errors in source assignments and, hence, fails to account for possible bias and uncertainty in the stock composition estimate.

Mixture modeling takes advantage of new knowledge about mixture composition as individuals are assigned to sources. Frequentist and Bayesian versions of this approach have been developed. In the frequentist approach, the conditional maximum likelihood estimate of stock composition is found by nonlinear search to maximize the probability of the natural marks occurring in the mixture sample when considered as a function of the unknown stock proportions. The conditional maximum likelihood method essentially allocates each mixture individual to the source stocks in proportion to its estimated posterior source probabilities, *i.e.*, the estimated fractions of individuals with the same natural mark in the mixture that are contributed by various source stocks. If necessary, individuals can be assigned as entities to the sources using the MAP rule. The conditional maximum likelihood estimate of the mixture composition equals the averaged allocated proportions among mixture individuals.

The Bayesian approach extends the mixture modeling approach to include estimation of both the stock composition and the allele relative frequencies in contributing stocks. A chain of assignments and mixture composition estimates are generated, in which mixture individuals are randomly assigned at each step in the chain to the source stocks with probabilities equal to their current estimated posterior source probabilities. Relative frequencies in the source baseline samples and the stock composition of the mixture are then updated from the most recent assignments of mixture individuals. Chain averages of the posterior source probabilities for individuals can be used with the MAP rule, if their assignments as entities are needed.

Empirical evaluations with individuals drawn from known populations indicate that the mixture modeling method performs considerably better at estimating stock proportions than does the classical individual assignments method. For example, classical individual assignments correctly identified only 15 of the 56 wild Atlantic salmon (27%) in a sample from a stock of a Scandinavian river, but the Bayesian posterior average was 91%, and the MAP rule (applied to the chain averages of posterior source probabilities for individuals) correctly identified 55 of the 56 individuals (98%).

Aggregating Chinook Stocks for Harvest Management and an Improved Genetic Stock Identification

Finding 19. *A comprehensive knowledge of all stocks is unnecessary if stocks can be aggregated into groups by assuming that stocks in a group share common characteristics that subject the stocks to the same or similar exploitation rates. Similar biology and recency of common ancestry, measured by genetic similarity, should govern how stocks are aggregated.*

Finding 20. *Aggregation schemes inconsistent with genetic relationships among stocks reduce the accuracy and precision of GSI, thereby limiting the usefulness of genetic analyses, and compromising the ability to manage fisheries with a full suite of data. The use of phylogenetic methods to identify genetically similar populations increases GSI accuracy. A Statistical Networks procedure (SN) was superior to two other aggregating procedures for identifying monophyletic groups of Chinook salmon populations in Puget Sound.*

Finding 21. *Standard quantitative stock aggregations should be designed coast-wide to be consistent with the phylogenetic relationships of stocks, and to maximize value to address specific fishery management needs.*

Harvest management of salmon requires information on temporal and spatial distributions, exploitation rates, escapements, spawner abundance, productivity, and basic biology of stocks. These variables are difficult to quantify for every stock potentially encountered in a fishery. A comprehensive knowledge of all stocks is unnecessary if stocks can be aggregated into groups by assuming that stocks in a group share common characteristics that subject the stocks to the same or similar exploitation rates. Similar biology and recency of common ancestry, measured by genetic similarity, should govern how stocks are aggregated. However, many aggregations used for management include stocks of similar geography, run-timing, and management activity, but not necessarily genetically related stocks.

As an alternative to stock groupings in TCT reports and NOAA Fisheries assessments, we designed a method based on Rannala and Mountain (1997) and illustrated it with data for 25 Chinook salmon stocks in Puget Sound and the Strait of Juan de Fuca (GAPS 2.1) and for additional samples collected by WDFW in the past year. First, we calculated the probability that a multilocus genotype occurred in each of the 25 stocks and averaged these probabilities for the 25 stocks. Second, these probabilities were randomized 10,000 times, and a new mean probability for each population was calculated after randomization. If the observed mean probability was equal to, or greater than, the 95th percentile of the randomized probabilities, we considered the observed mean probability to be significant. Third, we graphically joined populations that had significant mean probabilities into a network (Figure 4). This procedure revealed two large clusters of stocks connected to each other at two points: Lower Skagit–Samish rivers and Snoqualmie–Nisqually rivers. Nooksack River Spring and White River Spring Chinook salmon populations were wholly independent, and the Elwha and Dungeness river

populations were connected to each other, but not to other populations. Five stock groups could be distinguished within the two large clusters that were similar to the TRT groups.

We then constructed a phylogenetic tree using shared alleles at 13 microsatellite loci and neighbor-joining rooted by two Middle Fraser River stocks to search for monophyletic groups. Stocks within a monophyletic group are expected to have similar development, life histories, behaviors, and ocean distributions, and would therefore likely occur in a particular fishery. If true, management of monophyletic groups have greater predictive power than management of polyphyletic or paraphyletic groups. Hence, the use of monophyletic stock groups would be superior in a harvest management program. For the data here, none of the three aggregating procedures yielded monophyletic groups, but the Statistical Networks procedure (SN) was superior to the other two aggregating procedures. With both the TRT and SN procedures, the Puget Sound Spring/Summer group was paraphyletic with respect to the Skykomish, Snoqualmie, and Lower Skagit Fall runs. The TRT procedure also produced a paraphyletic Snohomish River group. All groups in the CTC procedure were paraphyletic, except for Hood Canal and for Strait of Juan de Fuca (which was not considered by the CTC).

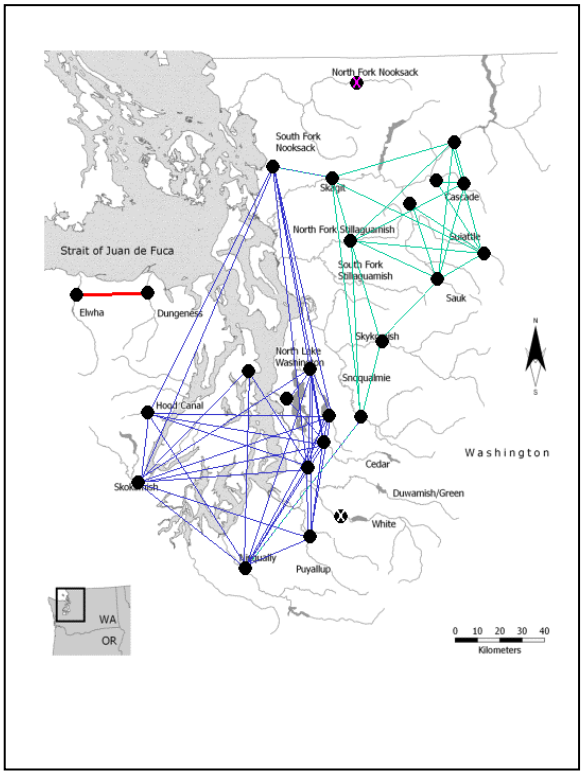

Figure 4. Statistical Network of Chinook salmon populations in Puget Sound.

GSI error rates for the management groupings generated by the three aggregating protocols were estimated with the CV-ML procedure (Anderson et al. unpublished). We used 100%-simulated mixtures to estimate error and pooled the stocks within a management group to obtain a single proportion estimate for the aggregation. This process was repeated 10,000 times to produce a

distribution of estimated proportions. The SN produced the lowest error rates and the CTC procedure the highest. The median value for four of the five management groups in the SN procedure was 1.00, while the value for the fifth group was 0.98. That is, for each group one-half of the 10,000 runs produced an error rate of 2% or less. The highest error rate was 58% for the Hood Canal group under the CTC procedure. Some outliers appeared for each grouping method and were due to inclusions of different life-history types in a population sample.

Fishery managers use stock composition estimates to assess catch allocation compliance and harvest impacts, which are measured on aggregates rather than on specific stocks (unless ESA issues are important). Although various rules can be used to aggregate stocks, these aggregations affect the efficacy of genetic stock identification (GSI). Aggregation schemes inconsistent with genetic relationships among stocks reduce the accuracy and precision of GSI, thereby limiting the usefulness of genetic analyses, and compromising the ability to manage fisheries with a full suite of data. Standard quantitative stock aggregations should be designed *coast-wide* to be consistent with the phylogenetic relationships of stocks, and to maximize value to address specific fishery management needs.

How different sources of error affect the accuracy of genetic stock identification

Finding 22. Several sources of error influence the accuracy of GSI estimation.

a. Sampling of the fishery. Error arises from small sample sizes and from non-random sampling.
b. Random sampling fails to include all stocks present in the fishery.
c. Sampling a finite number of genetic markers.
d. Genotyping error.
e. Errors on allele frequencies from sampling a finite number of individuals in baseline populations.
f. The inclusion of fish in the fishery sample from populations not in the population baseline also introduces error.

One important step in improving GSI accuracy is to identify the various sources of error. One source of error is from sampling of the fishery, but can be reduced by enlarging samples sizes, or by sampling larger time periods or more fishing boats. A second source of error arises even when the fishery is sampled randomly; the random sampling simply fails to include all stocks present in the fishery. Again, larger fishery sample sizes may reduce this source of error. A third source is due to the sampling of a finite number of loci. The inclusion of additional markers can potentially help to reduce this error. A fourth source of error is due to genotyping error in the laboratory. A fifth source arises from errors on allele frequencies from sampling a finite number of individuals in baseline populations. Lastly, the inclusion of fish in the fishery sample from populations not in the population baseline also introduces error. Here, three of these sources, fishery sample size, locus sampling, and baseline allele-frequency estimation are examined with a set of empirical data for a Chinook salmon fishery off southeastern Alaska.

Finding 23. *Partitioning of total expected square error (ESE)—a variance-like variable, including the effect of bias—into components b, c, and e showed that the largest source of error was due to uncertainties in allele frequencies in baseline populations (87.5%). A smaller proportion was due to fishery sampling (9.5%) and a very small proportion is due to genotypic sampling (2.7%). As the fishery used for this study was typical of other fisheries, these results likely show general trends for GSI estimates for other fisheries.*

A convenient measure of how much estimates are expected to be wrong is total expected square error (ESE). This is like a variance, but also includes the effect of bias. The goal is to partition ESE into three components. Calculating the portion of the ESE due to baseline deficiencies requires knowledge of baseline allele frequencies with certainty. However, the problem is that these allele frequencies are only based on estimates. One improvement is to adjust population allele frequencies to account for the increase in apparent divergence among populations due to finite sampling. Here, an unpublished algorithm was used to decompose the ESE into three sources of error in GSI estimates for the Chinook salmon fishery based on the GAPS population database. The largest source of error was due to uncertainties in allele frequencies in baseline populations (87.5%). A smaller proportion was due to fishery sampling (9.5%) and a very small proportion is due to genotypic sampling (2.7%). As this fishery was typical of other fisheries, these results likely show general trends for GSI estimates for other fisheries.

Finding 24. *When the level of differentiation among baseline populations is low ($F_{ST} < 0.01$), increased sampling will improve GSI accuracy. In other cases, statistical approaches can be used to improve baseline allele frequencies with two approaches.*

- *a. Generalized expectation maximum algorithm uses mixture samples to improve allele-frequencies estimates in the baseline populations. This approach is incorporated into available Bayesian GSI methods.*
- *b. The use of spatial models can be used to improve allele-frequency estimates by assuming that nearby populations tend to be similar.*

Further analyses show that when the level of differentiation among baseline populations is low ($F_{ST} < 0.01$), increased sampling will improve GSI accuracy. In other cases, statistical approaches can be used to improve baseline allele frequencies. One approach is to use a generalized expectation maximum algorithm that uses mixture samples to improve estimated allele frequencies in the baseline population. This approach is incorporated into available Bayesian GSI methods. Another approach might be to use spatial models to improve allele-frequency estimates by assuming that nearby populations tend to be similar. Models based on this assumption are used in epidemiology to map disease occurrence.

Finding 25. *The magnitude of error from unsampled source populations can potentially be estimated with three approaches.*

- *a. Simulations to examine the impact of excluding some existing populations from baselines.*

b. Spatially explicit models of population structure can be constructed to estimate allele frequencies of unsampled populations, and these estimates could be used in conventional GSI simulations.
c. Bayesian missing-data model may successfully identify unsampled populations contributing to a fishery.

Other problems were not addressed in the simulations presented here. The most vexing source of error is unsampled source populations. Three approaches could be used to evaluate the magnitude of this problem. First, simulations could be made to examine the impact of excluding some existing populations from baselines while keeping them in mixtures, to which GSI is applied. Second, spatially explicit models of population structure could be constructed to estimate allele frequencies of unsampled populations, and these estimates could be used in conventional GSI simulations. Third, a Bayesian missing-data model may successfully identify unsampled populations contributing to a fishery.

Finding 26. *Error in GSI estimation of low-contributing stocks in a mixture is difficult to evaluate, but can be reduced in part by larger fishery sample sizes. Current GSI algorithms tend to bias stock composition estimates toward 1/k, where k is the number of stocks contributing to the baseline.*

Another problem is error in the estimation of the frequencies of low-contributing stocks. If a stock of fish occurs at a low frequency in a fishery, a large fishery sample size is needed to accurately reflect the composition of the fishery. Current GSI algorithms tend to bias stock composition estimates toward 1/*k*, where *k* is the number of stocks contributing to the baseline. Hence, estimates of the frequency of a rare stock in a fishery will be biased upward. This bias is greatest when a rare stock is genetically similar to an abundant stock, because fish from the abundant stock are likely to be "mistaken" for fish from the rare stock; this will not be balanced by mistakes in the other direction, because there are fewer individuals of the rare stock to be misidentified.

Both simulations and empirical approaches can be useful, but recent experience has demonstrated the limitations of relying on simulation, which assumes the neutrality of alleles. In particular, the choice of high-graded markers showing large differences among populations can violate assumptions of neutrality, as these markers may be under the influence of natural selection, or may be neutral, but show larger than average differences between populations. Hence, the examination of empirical datasets may provide the best means of assessing power and of identifying components of GSI error.

Intra-annual and inter-annual variation in stock composition of the Queen Charlotte Island troll fishery 2002-2006

Finding 27. *GSI estimates provide an opportunity to understand important features of salmon migration and spawning biology. In-season comparisons of GSI estimates in an area off northern British Columbia revealed contrasting abundance trends for*

fish from different areas. These estimates show shifts in stock compositions of fish in the troll fishery during the fishing season, which likely reflect migration patterns of various stock groups past the Queen Charlotte Islands.

Finding 28. *Annual comparisons of five major stocks in either the troll or commercial catches showed shifts for some regions but not for others. Annual variation was most pronounced for Oregon fish with the highest proportions in August 2004. Fraser River fish also increased in 2002 and 2006, likely reflecting the strong returns to the Thompson River drainage in those years.*

Finding 29. *Seasonal and annual comparisons are possible only after a large-scale population baseline has been established to identify stocks potentially contributing to a fishery. Seasonal and annual comparisons require frequent sampling during a fishing season to provide an accurate view of changes in contributing stocks.*

Results from mixed-stock analysis of Chinook salmon harvested off northern British Columbia provide an opportunity to understand important features of salmon migration and spawning biology. Stock compositions for Chinook salmon in either test troll fisheries or commercial troll fisheries off the northwest coast of the Queen Charlotte Islands from 2002 to 2006 varied over the monthly sampling cycle from April through September.

In-season comparisons of these samples revealed contrasting abundance trends for fish from different areas (Figure 5). The abundances of fish from Washington and Oregon progressively increased during the season from 4-6% in April to 21-58% in September. Chinook salmon from California followed a similar trend, but with much lower abundances. Fish from the Columbia River, however, were most abundant early in April (44%) and least abundant in September (11%). In contrast to these early or late proportions, fish from the Fraser River were prevalent mid season in May-July (27-36%). Fish from the east coast of Vancouver Island comprised a minor component of the fishery (1-4% monthly). However, fish from the west coast of Vancouver Island (WCVI) peaked at 19% in April and declined to 3% in September. The proportion of fish from northern British Columbia was highest in April (7%), and declined gradually to a low of 1% in September. These analyses clearly indicate that stock compositions of fish in the troll fishery change during the course of the season and likely reflect the migration patterns of various stock groups past the Queen Charlotte Islands.

Inter-annual comparisons of the proportions of five major stocks in either the troll or commercial catches showed shifts for some regions but not for others. Inter-annual variation was most pronounced for Oregon fish in August (15-45%), with the highest proportions in August 2004. Contributions from the Columbia River and Washington were relatively stable among years. Fraser River fish displayed high proportions in all months during 2002 and 2006, likely reflecting the strong returns to the Thompson River drainage in those years. In most months, the proportion of WCVI fish was small relative to other major stocks, and thus the absolute level of annual variation for this stock was less in comparison the other stocks.

The comparisons presented here were possible only after a large-scale population baseline was established so that all the stocks potentially contributing to a fishery could be identified. While

proportion estimates are an important starting point, abundance data or additional sampling may be required to extrapolate the results of a comparison such as this to other regions or fisheries. Abundance data are also required to refine inferences of distribution and migration patterns. An important result of these in-season and inter-annual comparisons is that reasonably frequent sampling during a fishing season is required to provide an accurate view of the presences of various stocks contributing to a fishery.

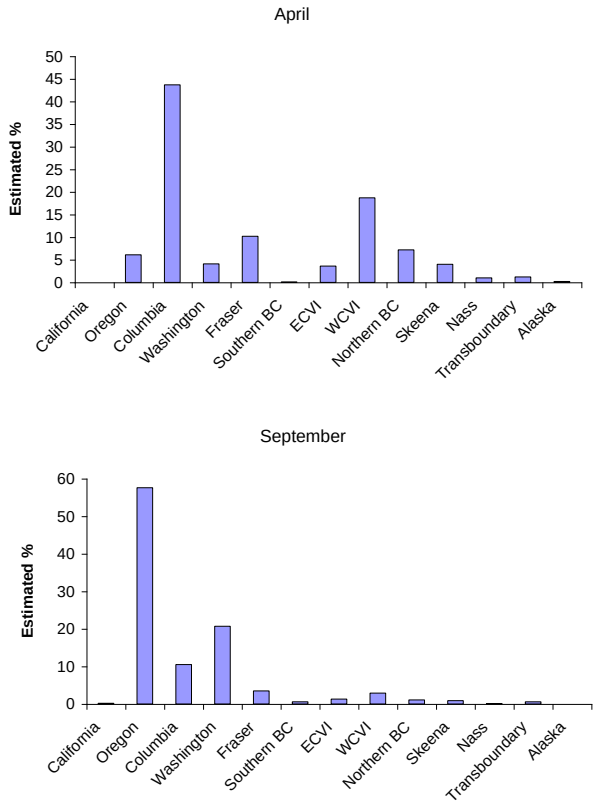

Figure 5. GSI estimates of population origins for April and September in northern British Columbia Chinook fishery.

MAJOR RECOMMENDATIONS

Recommendation 1. *Support continued development of genetic markers (particularly for SNPs in sockeye salmon coast-wide)*

a. *Use appropriate lessons from the GAPS approach to marker standardization for the development of population baselines for additional species.*
b. *Develop appropriate markers for use in a coast-wide baseline.*

Genetic markers have become firmly established in the toolbox of methods used in the fishery management of Pacific salmon. As new markers and statistical methods become available they should be incorporated into this toolbox, when they improve population resolution, when they can be genotyped with high-throughput methods, and when they are suitable for addressing broader problems in fishery management.

New markers and methods of genotypic analysis must be standardized among laboratories to support the development of coast-wide data baselines for the various species of Pacific salmon. Until recently, microsatellites have been the marker of choice of most laboratories and great effort has been made to standardize loci and alleles. This standardization process in Chinook salmon has produced the large GAPS baseline that is used in many regional and coast-wide applications, including but not limited to GSI. Many of the lessons learned from the creation of the GAPS baseline can be applied to the development of baselines for other species. One is the use of allelic ladders to standardize genotyping among laboratories.

The standardization of SNP markers, on the other hand, is easier because SNP variants are defined by base changes at a specific nucleotide site in a DNA sequence that can be easily recognized among laboratories. Nevertheless, discrepancies among laboratories may still arise from handling errors, or from the use of different assay chemistries. Standardization of either marker type is relatively simple, but it is not yet clear in practical application how standardization costs will compare for the hundreds of SNPs that will likely be required to compare to coast-wide microsatellite datasets.

Recommendation 2. *Maintain and improve existing standardized microsatellite population baselines*

a. *Existing microsatellite baselines provide the only means of addressing some management problems.*
a. *These baselines should be maintained and extended to provide greater levels of population resolution.*

Microsatellites are the current interagency standard for use in a broad range of applications in population and ecological genetics of Pacific salmon. Even though SNPs hold considerable promise, especially in specific GSI applications, the current capacity for addressing problems with microsatellites should be maintained until coast-wide comparisons with SNP markers have

been made to decide future directions. Both microsatellites and SNPs appear to provide similar levels of population resolution and both classes of marker are capable of supporting the development of numerous additional markers. However, at the moment several on-going applications are possible only with microsatellites, because the utility of SNPs, in especially coast-wide applications, has not yet been demonstrated.

Recommendation 3. *Empirical comparisons of SNPs & microsatellites on a coast-wide scale, with focus on Chinook and sockeye.*

a. Even though SNPs often provide a high level of resolution for discriminating among regional populations, can they be effect in a coast-wide baseline?
b. The particular sample of SNP or microsatellite loci for a regional comparison can determine the outcome of a comparison. Hence, appropriate marker should be used in a comparison.
c. Simulations can be used to assess the level of resolution that a marker provides to discriminate among a group of populations.
d. Blind samples of known origins should be used in a GSI analysis to examine resolution of marker types.
e. *Evaluations between marker types should be posed in terms of cost for a given amount of population resolution, not just the cost of genotyping.*

Members of the Genetics Workgroup agreed that the most urgent recommendation was an empirical evaluation of microsatellites and SNPs in a coast-wide setting. The question of resolution and numbers of SNPs required for coast-wide applications are best determined empirically with direct comparisons of SNP and microsatellite baselines. To date, this analysis has not been possible because SNP data are largely local or regional. The only coast-wide SNP data sets are still sparse both in terms of the number of lineages, populations, and individuals.

Some in the Pacific salmon research community believe that SNPs will eventually replace microsatellites in most applications, including GSI. At this point, however, more SNPs—perhaps many more—will be needed to provide the broad utility now enjoyed by microsatellites. The hope by some is that this transition would proceed rapidly, but experience is showing that the development of SNP markers and assays is more expensive and time consuming than initially envisioned. Without a substantial increase in funding, it is likely to be some years before SNPs reach widespread implementation, especially in descriptive population genetics and a wide range of non-GSI conservation and restoration applications. SNPs will undoubtedly play an increasingly larger role in salmon GSI applications, although the timing of that transition remains unclear.

These urgently needed comparisons should also account for ascertainment bias in the selection of, especially SNP, markers. These high-graded markers have been highlighted as a limitation of SNPs in some population genetic studies. Also required in these comparisons is a cost-benefit analysis of the amount of population resolution provided for loci-allele combinations. Despite the optimism that the cost of developing SNP markers would be comparable to cost of developing microsatellites markers, the cost of SNP marker development for most salmon species has been substantially higher than that for microsatellite markers.

Recommendation 4. *The potential of a marker type to resolve features of salmon population dynamics in addition to GSI (mixed stock analysis) should be considered before adopting the marker.*

a. Most models of population structure assume the selective neutrality of alleles.

b. High-graded markers showing strong differences among populations may improve GSI estimation, but produce biased estimates of demographic parameters.

In theory, both microsatellites and SNPs should work equally well for GSI, and with avoidance of ascertainment bias, perhaps all applications. In practice, however, there is a danger that the development of only high-graded markers for GSI applications may greatly limit the traditional use of genetic markers for a variety of other applications in population genetics, and conservation. At present, a greater number of population problems can be addressed by microsatellite baselines than by SNP markers, because of the greater availability of microsatellite markers and the large coast-wide baseline data sets.

Recommendation 5. *Support studies investigating sources of GSI error. Preliminary results of theoretical and simulation studies point to ways to improve GSI accuracy.*

a. Investigate ways of improving allele-frequency estimates of populations in baseline. Only marginal gains in accuracy can be achieved with larger samples of fishery mixtures and genetic markers.

b. Support studies of other sources of GSI error, including upward bias of low-frequency stocks in mixture, and missing baseline populations.

c. *Adopt* mixture modeling *for GSI estimation.*

Improvements in statistical methods for boosting the accuracies of GSI estimates are far less costly than the developments of new high-resolution molecular markers. New markers may also require modifications of established methods. Hence, studies of statistical methods should be encouraged. One important area concerns the best algorithms for making GSI estimates. The statistical treatise presented in Appendix D concluded that mixture modeling and not individual assignments and summing provided greater GSI accuracy. Strides have also been made by using Bayesian methods to estimate stock compositions and to improve baseline allele frequencies. Errors in baseline allele-frequency estimation were identified in an analysis presented in Appendix F as an important source of GSI error. Additional GSI problems await more in-depth statistical treatment. One important problem is the upward bias in estimates of low-frequency stocks on a fishery sample.

Recommendation 6. *Re-examine methods used to aggregate baseline stocks into reporting groups to increase GSI accuracy.*

In the same vein as the previous recommendation, improvements in stock aggregation methods can greatly improve GSI accuracy with minimal costs. In the absence of a comprehensive knowledge of stock attributes, stock aggregation can be used to group populations with common characteristics that subject them to the same or similar exploitation rates. Similar biology and recency of common ancestry, measured by genetic similarity, should govern how stocks are aggregated. Presently, many aggregations used for management include stocks of similar geography, run-timing, and management activity, but not necessarily genetically related stocks.

Recommendation 7. *Support summary studies of seasonal and multi-year GSI results to better understand the ocean biology of Pacific salmon.*

Although GSI estimates in mixed-stock fisheries can provide in-season information to manage escapement, collections of GSI estimates for a fishery during a season or among years can provide insights into the ocean biology of particular stocks. Several datasets are now available that can be used to make these inferences, and the PSC should encourage the syntheses of these datasets into a broader coast-wide picture of Pacific salmon ocean migration pathways and abundances.

Recommendation 8. *Support collaborations between geneticists and population modelers and harvest managers to enhance the utility of GSI results.*

Many issues were not addressed in the time available to the Genetics Workgroup. The overall goal of the workshops was to explore ways that GSI could be better incorporated into fishery management. However, the focus of the WG was largely on the development of new molecular markers and on improvements in statistical procedures, and not on the broader issues of modeling and management. While the two workshops provided opportunities for geneticists, statisticians, modelers, and managers to exchange views, a wider synthesis of genetics into these other fields still remains to be achieved. The summaries and views presented in this report represent a start toward this synthesis.

PART IV. APPENDICES

APPENDIX A. CHOICE OF MARKER TYPES FOR GENETIC STOCK IDENTIFICATION

APPENDIX A. Choice of Marker Types for Genetic Stock Identification[1, 2]

Christian Smith
Abernathy Fish Technology Center, US Fish and Wildlife Service, 1440 Abernathy Creek Road, Longview, WA 98632

Kristi Miller, Terry Beacham
Pacific Biological Station, Department of Fisheries and Oceans, Nanaimo, BC, V9R 5K6

Chris Habicht
Gene Conservation Laboratory, Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, AK 99518

Lisa Seeb
School of Fisheries, University of Washington, Seattle, WA 98195

Steve Latham
Pacific Salmon Commission, 600-1155 Robson Street, Vancouver, BC, V6E 1B5

Paul Moran
Conservation Biology Division, NOAA Fisheries, 2725 Montlake Boulevard East, Seattle, WA 98112

W. Stewart Grant
Department of Biological Sciences, University of Alaska Anchorage, Anchorage, AK 99506

---

[1]While every attempt was made to produce a consensus view of marker development and future applications, the conclusions expressed in this document may not represent the views of all authors.

[2]'Genetic stock identification' (GSI) is a broad concept, including both the identification of genetically discrete populations and 'mixed stock analysis' (MSA). In common usage, however, GSI has become synonymous with MSA.

INTRODUCTION

New technologies periodically appear and should be considered in applications of genetic methods for the management of salmon species of interest to the Commission. Allozymes were replaced with microsatellite as a marker of choice in salmon population studies, because microsatellites offered several advantages (see Box A1, *e.g.* Small et al. 1998; Beacham *et al*. 2001). A new technology, single nucleotide polymorphisms (SNPs), has been developed for several applications in genomic and population research and has recently been applied to resolving management problems in Pacific salmon (*e.g.* Smith et al. 2005b). The goal of this section is to discuss the relative merits of these molecular markers for GSI, but this discussion has to be placed in a broader perspective than just the focus on a single application.

A new population marker should possess three characteristics: 1) equal or greater resolution of population differences than for existing markers, 2) high throughput genotyping for applications often requiring the analysis of thousands of fish annually, and 3) suitability to continue a well-established tradition of research on salmon population biology. A cost-benefit analysis of these factors is needed before a new marker can displace previous markers and be adopted for general use. The core use of molecular genetic markers in Pacific salmon has been to describe various aspects of genetic population structure; that is, to estimate the degree of genetic connectivity among populations, inbreeding, migration and effective population size, among other variables. An extension of this has been the use of molecular markers to make individual assignments to parents or populations, or to estimate stock proportions in mixed-stock harvests, and this latter application is the focus of this document (Table A1).

A large toolbox of population genetic models can be used to interpret genotypic data. Genomic markers have provided novel insights into numerous kinds of demographic events in Pacific salmon, including the estimation of genetically effective migration rates (Grant 1997) and population sizes (Waples 1990), inbreeding and outbreeding, and historical founding events (e.g. Teel *et al*. 2001; Beacham *et al*. 2003). Most models used to make these kinds of population inferences assume that the markers are neutral to natural selection. Markers showing biased allele-frequency differences among populations because of regional selection are unsuitable for these models. Hence, some loci are more useful than others to address these research problems. While several classes of genetic markers, including allozymes, denaturing gel gradient electrophoresis (DGGE; Fischer and Lerman 1983), and amplified fragment length polymorphism (AFLP; Vos *et al*. 1995), have been useful for some GSI applications (*e.g.* Beacham *et al*. 2005; Flannery *et al*. 2007), these markers have had limited in their flexibility for addressing a wide gamut of problems.

One important distinction among markers is whether a technology defining a marker type *surveys* genetic variability at a locus or *assays* a predetermined polymorphism. Polymorphism assays, while useful for some population applications and mixed stock analysis, are limited in their use to measure levels of genetic diversity. The use of genetic survey markers has been instrumental for detecting loss of genetic diversity through poor hatchery practices (Allendorf and Phelps 1980; Ryman and Ståhl 1980; Busack and Currens 1995) or through founder events and population bottlenecks (Luikart *et al*. 1998; Garza and Williamson 2001).

Information Box 1. *Genetic markers*

Genetic markers reflect different classes of genetic variability. For example, allozyme markers reflect non-synonymous coding changes in DNA that produce differences in size or charge of a protein product. These two properties can facilitate electrophoretic separation in a supporting medium. Microsatellite markers are based on changes in the number of tandem repeats. The insertion or deletion of a repeat motif can be detected with electrophoresis. Single nucleotide polymorphisms (SNPs) are single-base differences assayed by interrogation of a single nucleotide position in a DNA sequence. DNA sequence polymorphisms among individuals provide the basis for genetic assignments. The portion of a DNA sequence that is polymorphic among the taxa of interest is called a "genetic marker". Rapid and inexpensive assays have been developed to allow the inference of either the DNA sequence or some property of the DNA marker.

Microsatellites have been used over the last decade for PSC-related GSI applications (*e.g.* Small *et al*. 1998; Beacham *et al*. 2001; Beacham *et al*. 2004b; Beacham *et al*. 2007a) and are the current interagency standard for a broad range of applications in population and ecological genetics of Pacific salmon. However, SNPs hold considerable promise, especially for specific GSI applications. Although SNP assays (largely with restrictions enzymes) were available before the development of microsatellite methods (Botstein *et al*. 1980; Moran *et al*. 1997), the lack of high-throughput assays made SNPs less appealing than allozymes and microsatellites. The development of novel chemistries facilitating high-throughput genotyping (Kwok 2003) has stimulated renewed interest in SNPs (*e.g.* Smith *et al*. 2005b).

BACKGROUND

The first SNPs for fishes were developed in model species (rainbow trout and Atlantic salmon) to conduct genome-wide screens for quantitative trait loci. Large numbers of SNPs (1000's to 10's of 1000's) have been surveyed in relatively small numbers of individuals to assess linkage with phenotypic traits of interest. SNPs are gaining popularity in population genetics studies, particularly because they offer promise in resolving adaptive variation among populations (Lui *et al*. 2005). SNPs have also recently been used for individual identifications (Seddon *et al*. 2005), pedigree analysis (Werner *et al*. 2003), and cultivar selection for breeding programs (Shirasawa *et al*. 2006).

The large-scale use of SNPs for population studies is new and the salmon genetics community is at the forefront in the use of SNPs for mixed-stock analysis in harvest samples. Although microsatellites are the current standard for general molecular genetic research on Pacific salmon, some researchers in the salmon research community believe that SNPs may replace microsatellites in many applications including GSI. More SNPs—perhaps many more—will be needed to provide the broad utility now provided by microsatellites. Once a SNPs population baseline is established, a subset of these SNPs can be used for particular applications (see Lui *et*

al. 2005). The utility of SNPs for descriptive population genetics, restoration and conservation applications remains to be demonstrated. Some hoped the transition to SNPs would be rapid, but experience is showing that it is more expensive and time consuming to develop robust SNP assays suitable for coast-wide applications than to develop microsatellites. Without substantial funding the widespread implementation of SNPs is likely to be some years away.

DIFFERENCES BETWEEN SNPs AND MICROSATELLITES

Resolving power

Several factors potentially influence the level of resolution achievable with a molecular marker. Theoretical results show that for markers uninfluenced by natural selection, the resolution of population differences (Ryman *et al*. 2006) or of populations in a mixed fishery sample (Kalinowski 2004) depend on the number of independent alleles at loci. The number of independent alleles for a locus is $r - 1$, where r is the number of alleles segregating at the locus. Consequently, a single SNP locus, if assayed for only one nucleotide change, has one independent allele, whereas a highly polymorphic microsatellite locus can have more than 50 or more independent alleles.

The independent-allele rule, however, fails to capture the interaction between the numbers of loci and alleles in empirical baselines in providing statistical power. A simulation study of assignments of individuals to parents found that adding alleles and loci interactively improved assignments (Bernatchez and Duchesne 2000). The success in allocating individuals to populations, on the other hand, was more influenced by an increase in the number of loci, but for a given number of loci, gains in success were achieved by including more alleles. In the Bernatchez and Duchesne model, moderately polymorphic loci with 6–10 alleles appeared to provide the best allocation success.

However, empirical evaluations for sockeye salmon showed that loci with larger numbers of alleles provide greater resolution among populations (Beacham *et al*. 2005). Loci with large numbers of alleles also provide greater resolution than less polymorphic loci among Chinook salmon populations regionally (Beacham *et al*. 2007a, b) and across the North Pacific (Beacham *et al*. 2006a, b). Loci with 6-10 alleles were among the poorest performers for discriminating among populations. In the latter study (Beacham *et al*. 2007b), the resolving power of 9 SNPs was similar to that of a single microsatellite locus with 17-22 alleles. Geographically large-scale comparisons between microsatellites and SNPs remain to be made.

Another factor influencing power is the interaction among the number of alleles at a locus, samples size, and the accuracy of frequency estimation. As the number of alleles per locus increases, so should the number of individuals in the baseline. For example, in a SNP sample of 100 alleles (50 fish), high- to moderate-frequency alleles are estimated with some confidence with an error given by multinomial sampling theory. For highly polymorphic microsatellite loci (*e.g.* 50 alleles), a population sample of 50 fish produces a proportionately larger error on allele frequencies. Rare alleles often remain unsampled.

Table 1. Characteristics of molecular marker used in fishery management

Characteristic or use	**Microsatellites**	**Single nucleotide polymorphisms** (Nuclear loci)
Statistical power	Highly polymorphic loci provide the most power per locus for detecting differences between populations	Biallelic SNP loci have less power per locus than highly polymorphic microsatellite loci. Assuming selective neutrality and random selection of loci (no ascertainment bias), the number of alleles roughly corresponds to statistical power.
Marker development	Moderate cost. GenBank sequences available and cross-species PCR amplifications often possible	Presently, screening for polymorphisms is somewhat costly, but costs are expected to drop with development of additional discovery technologies. Cross-species SNPs assays generally not possible. Genomic duplications complicate SNP development
Routine genotyping	Moderate costs. Multiplexing of several loci possible. Bulk runs bring down costs	Multiplexing in development with the promise of low per-locus costs with bulk analysis. Biotech or core-lab genotyping of SNPs possible.
Parental assignment	Large amounts of statistical power	Large amounts of statistical certainty when numbers of when large numbers of SNPs markers are used
Mixed-stock analysis	Large amount of statistical power for regional and coast-wide GSI. Depends on level of divergence among populations	Large amount of statistical power demonstrated for regional analyses. Depends on level of divergence. Coast-wide power of regionally developed SNPs not tested.
Within-population genetic diversity	Relative comparisons can be made among samples. Surveys existing diversity.	Affected by choice of SNPs. Assays a predetermined polymorphism, and marker frequencies may be influenced by selection.
Inbreeding	Inbreeding indices, heterozygote deficit	Inbreeding indices can be used, but loss of information because only two allelic states are assayed
Detection of outbreeding or hybridization	Heterozygote excess, hybrid indices	Heterozygote excess, hybrid indices, but information content is low because of only two

		alleles. However, the reduced numbers of alleles may be partially offset by larger numbers of loci.
Gene flow (‘straying’)	Possible with numerous models when alleles can be assumed to be neutral	Natural selection or ascertainment bias more likely to violate assumptions. Statistical tests for selection can be used to identify neutral alleles.
Effective population size	Models assume neutrality	Assumption of neutrality may be violated, but tests for selection can be used to identify neutral alleles.

In other applications, a recent study of over 15,840 SNPs (on a SNP array) and 328 microsatellite loci in humans showed that the information content of random microsatellites was four to twelve times greater, on average, than that of a randomly chosen SNP (Lui *et al*. 2005). However, some SNPs were more informative than single microsatellites and this finding suggests that highly informative SNPs can reduce the number needed to match the resolution of microsatellite baselines. Unfortunately, large numbers of SNPs are still unavailable, except for Atlantic salmon and Rainbow trout, which have been the focus of large scale sequencing studies. This deficit could possibly be circumvented by using highly informative SNPs or by focusing on genes thought to be of adaptive significance (*e.g.* QTLs from other species).

One key difference between microsatellite loci and SNPs is that, unlike microsatellites, SNPs may not be polymorphic coast-wide. While ascertainment bias (Box 2) can be used to identify highly informative SNPs for a particular region, a different set of informative SNPs may be required for other regions or for coast-wide baselines. The most effective, high-resolution baselines will likely contain both adaptive and neutral loci, with neutral loci providing strong regional resolution, and adaptive loci identifying particular local populations. The question of resolution and numbers of SNPs required for coast-wide applications will ultimately be determined with empirical comparisons of the effectiveness of SNP and microsatellite baselines.

Information Box 2. Ascertainment bias

The use of highly informative SNPs (ascertainment bias) aids in the choice of markers for resolving allele-frequency differences among populations. Ascertainment bias can be advantageous for GSI, but detrimental for other applications in population genetics, conservation and evolutionary systematics. However, this result is not limited to SNPs, as highly informative microsatellite loci can also be identified.

The better performance of some loci relative to others may be due to two sources:

1). Some alleles show greater than average differences by chance from reproductive sampling each generation. This increased resolving power is not due to natural selection.

2) Greater resolution of some loci may reflect directional natural selection (Ford 2002; Schlötterer 2002). Usually a greater number of SNPs are surveyed to achieve the same level of resolution as with microsatellite markers. Hence, a grater number of SNPs, selected because of their resolving power, may be influenced by selection.

One possible drawback of using markers under selection is that baseline allele frequencies may be unstable during times of rapid climate change, so that periodic surveys may be important. For many applications, the effects of natural selection are immaterial, but for others, such as determining the demographic histories of populations, they are problematic.

While mitochondrial (mt) DNA sequences are not currently being used to survey variability among salmon populations, mtDNA variants occur in the repertoire of SNP markers. These organellar DNA variants are expected to show different patterns of variability from nuclear variants, because they are maternally inherited, usually without recombination at replication, and occur in an individual as a single haploid copy. This mode of inheritance confers an effective population size that is about one quarter that of nuclear diploid markers, and hence is subject to greater levels of random drift among populations. This expected high level of divergence between populations makes them attractive as population markers.

Marker development and throughput: Analysis of cost and time

In a comparison between marker types, several variables in addition to genotyping costs should be considered. Although costs per genotype are sometimes used to compare techniques, the cost per fish for a given level of resolution may be a better metric for overall comparison. Before a new marker technology can be adopted coast-wide, costs of replacement must also be considered. Replacement entails effort to establish and standardize new population baselines and to implement new infrastructure to provide real time estimates. In some cases, existing instruments can be used for both microsatellites and SNP analyses. Alternatively, SNP markers could be added to microsatellite markers to resolve particular problems not resolved with microsatellites. Dual laboratory capabilities, however, may be inefficient because of the costs of additional equipment and personnel training. Differences in cost between SNPs and microsatellites can be broadly described under marker development and routine genotyping.

Marker development– Primers for microsatellite markers developed for one species often work well on related species. Thousands of microsatellites have been isolated in various species of salmon and serve as a starting point for developing new microsatellite markers in other species. Hence, when a collection of microsatellite markers cannot resolve a particular problem, additional microsatellite markers can be developed rapidly and inexpensively from microsatellite sequences in GenBank (National Institute of Health DNA sequence repository).

In contrast, SNPs are not usually transferable among species, but must be developed anew for each species. Until recently, the easiest, most cost effective way to obtain SNPs is to search for polymorphisms in the EST (express sequence tags) and DNA sequence databases of GenBank. Unfortunately these sequences are largely limited to rainbow trout and Atlantic salmon. In the absence of these 'head-start' sequences, SNPs must be developed one at a time, and this development has been costly and time consuming. New-generation technologies that facilitate the rapid sequencing of long stretches of DNA may shorten developmental times and costs. Presently, numerous SNP markers have been developed largely for Chinook, chum, and sockeye salmon and a few markers for coho salmon (see section on available databases).

Marker genotyping–Both microsatellite and SNP methods use primer-defined polymerase chain reaction (PCR) amplifications of particular regions of a DNA sequence. One limiting step for both methods is sample dissection and DNA extraction. Some extraction methods produce DNA that can be archived longer than DNA extracted with other methods. While some methods of extraction produce DNA more rapidly, usually methods are used that produce high-quality DNA extracts with long storage lives. The development of robotic dissection and extraction procedures can improve turnaround times for in-season analyses for both methods.

A core set of microsatellite loci has been standardized in all laboratories associated with the PSC. Generally, polymerase chain reaction (PCR) is used to amplify a section of DNA containing the microsatellite region with standard PCR primers. This is followed by size fractionation of the PCR products with an automated DNA sequencer. PCR multiplexing (the amplification of more than one microsatellite locus at once) increases the efficiency of the microsatellite PCR reactions, but can lead to false peaks that can be mistaken for an allele. Another potential problem is that some alleles are preferentially amplified over others in a heterozygous genotype.

Genotyping costs vary from one laboratory to another and depend not only on the costs of materials, labor, and genotyping hardware (*e.g.* thermocyclers, DNA sequencers, robotic pipettes), but also on institutional requirements for cost recovery. Even so, costs for microsatellite genotyping appear to be similar among laboratories (within a factor of two or three). SNP genotypes can be assayed by a variety of methods. Most salmon fishery laboratories presently use the 5'-nuclease reaction implemented with TaqMan. The cost of TaqMan genotyping can vary by nearly an order of magnitude, depending on the costs of genotyping hardware, and the number of SNPs assayed. Thus, the number of SNPs that can be genotyped for a comparable cost of genotyping microsatellites varies widely among laboratories (Table 2). Estimating costs is further complicated because SNP and microsatellite loci contain different amounts of information depending on the particular application to population genetics or fishery management.

Table 2. Platforms presently used by laboratories for SNP genotyping with TaqMan for the PSC. 'Number of SNPs to run' indicates the number of SNPs that can be genotyped for the same cost as a typical microsatellite panel

Infrastructure	Reaction volume	Number of SNPs to run
96-well reader	10-15 µl	18
384-well reader + robotics	5µl	42
Fluidigm	Nanolitre	87

Opportunities for automated genotyping are greater for SNP analyses. Most genotyping errors result from human-induced error and not from PCR amplification or instrumental errors. Hence, genotyping methods requiring a greater number of steps by technicians may be more prone to error than automated methods, especially methods requiring repetitive procedures susceptible to technician fatigue. Presently, microsatellite genotyping requires about twice the handling of a PCR product than does SNP genotyping. However, a microsatellite locus contains, on average, more than twice as much information as a SNP, so that microsatellites and SNPs may have

similar levels of experiment-wise errors. Costs may be reduced with the acquisition of multiplex technology or by outsourcing genotyping to a core agency laboratory or commercial laboratory.

Microsatellite scoring requires a greater number of visual inspections of computer images than does SNP genotyping. For example, in the analysis of 12 microsatellites in 10,000 fish, a technician is required to individually assess 12 x 10,000 = 120,000 images. In laboratories that double score for quality control, two technicians may spend a few days assessing genotypes on a computer screen. In comparison, a technician scoring 77 SNPs in 10,000 fish would individually assess either 77 x (10,000/384) = 2,079 images or just 77 images, depending on software. Again, this number will double in laboratories double-scoring for quality control. In assays of a few hundred individuals, the difference in effort between microsatellite and SNP genotyping may not be substantial, so that error due to technician fatigue may be insignificant. However, in assays of thousands or tens of thousands of fish, SNP automation can give an advantage in the time required to complete the scoring of genotypes and in error rate reduction. Although per locus error rates and throughput are favorable for SNPs, the net effect of lower error rates for very large numbers of SNPs remains uncertain.

In-season Mixed Stock Analysis

In-season fishery management in some areas has been guided by mixed stock estimates in either test or commercial fishery catches. Allozymes, microsatellites and, recently, SNPs have been used for in-season management, which requires rapid laboratory and statistical analyses of as many as 1000 fish in a day or so. While not all applications require this level of expediency, real-time GSI may become increasingly more important, as migration patterns of many stocks shift annually (Winther and Beacham 2006; Beacham Workshop report). Examples of in-season mixed stock identification appear in Box 3.

Several factors influence the ability of a marker to facilitate rapid turnaround times, including sample preparation (protein or DNA extraction), genotyping, data collection, and data analysis. Rapid SNP analysis for real-time applications may be somewhat limited. Unless a laboratory has invested in microarray technology, the number of SNPs that can be surveyed rapidly depends on the number of thermocyclers available for PCR, because Taqman assays survey one SNP at a time. Rapid throughput of large sample sizes depends on making hundreds of PCR reactions. Considerably more efficient genotyping platforms are required for SNPs to achieve similar turn-around times that are possible with microsatellite loci.

Standardization of data across agencies

Standardization of methods and datasets among laboratories is important for the development of a coast-wide data baseline for a particular species of salmon. The goal of standardization is to generate the same set of genotypes from the same samples in different laboratories, and to generate data in different laboratories that can be combined into a single dataset. The various steps for standardizing, based on the experience with GAPS, are outlined in another workgroup report.

Microsatellite genotypes consist of relative fragment mobilities, which often vary from one laboratory to another because of differences in genotyping platforms. Thus, reproducibility among laboratories requires that allele sizing be adjusted though laboratory standardization before new laboratories can add data to a standardized database or use those databases for mixed stock applications. Coast-wide standardizations include the use of a common set of loci, the exchange of tissues or alleles ladders (*e.g.* LaHood *et al*. 2002), allele curation and periodic testing.

Information Box 3. Use of mixed-stock analysis for in-season management

DFO: Fraser River sockeye salmon

While most Fraser River sockeye salmon populations are abundant and support an offshore fishery, some populations have been listed as 'endangered' by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC 2002). Fishery managers were concerned that late-run spawners returning to an endangered population in Lake Cultus would be vulnerable in the fishery. Earlier than expected returns of late-run fish led to high levels of mortality enroute to spawning areas. Over two months 9300 returning fish were genotyped for 14 microsatellite loci and one MHC locus (Beacham *et al*. 2004b). Samples including as many as 600 fish were delivered to the laboratory several times a week and results were returned to fishery managers within 9–30 hr. These results showed that the initial river entry of late-run Cultus Lake fish had been advanced by over 6 weeks. A large pulse of fish entered the fishery in mid August. These returns overlapped with Summer-run fish and precluded harvests of exclusively Summer-run sockeye. The estimated exploitation rate of Late-run Fraser River fish was 13% and fell within the management objective of 15%. The DFO maintains microsatellite baselines for sockeye, Chinook, coho, chum, and pink salmon, which have been used for GSI over the past ten years. In the past four years alone, 20,000 fish annually have been analyzed rapidly and used for "real-time" in-season management of harvests.

ADFG: Bristol Bay sockeye salmon

Over the last two years, ADFG genetics laboratory has used SNP analysis in real-time mixed stock analysis to help in the management of the Bristol Bay sockeye salmon fishery. Approximately 300 sockeye were analyzed every two days for about a month. As many as 390 fish can be genotyped for 44 SNPs in 16-20 hrs. New technology has enabled the ADFG to increase its throughput such that one technician can comfortably screen 20,000 genotypes in about 16 hrs (one day to run the samples, a morning to score the samples, and an afternoon to produce mixed stock estimates). New technology is being implemented that can assay 48 fish for 48 genotypes in one thermal cycler run lasting about 2 hr. If 48 SNPs are used in the analysis, about 400 fish (8 gene chips) can be genotyped in a short amount of time with one thermocycler. As with microsatellites, the time needed to dissect samples manually and to extract DNA limits rapid analyses. The time for genotype analysis may be further reduced by automated methods of DNA extraction.

Microsatellite data for Chinook salmon have been standardized among ten laboratories (http://www.nwfsc.noaa.gov/research/divisions/cbd/documents/gaps_year2_final.pdf) (Seeb *et al* submitted), for coho salmon between some laboratories (WDFW+CDFO, NOAA Seattle+USFWS Longview), and for chum salmon (DFO+USFWS Anchorage; DFO+WDFW, WDFW+NOAA, Seattle, DFO+University of Alaska, Juneau). The cost of microsatellite standardization has declined substantially and has become simpler and more robust with the use of allelic ladders (LaHood *et al*. 2002). Standardization of either marker type is now relatively simple, but in practice, the cost of standardizing hundreds SNP markers for coast-wide applications is yet uncertain.

DISCUSSION

The goal of this report was to evaluate the relative merits of microsatellites and SNPs for use in ocean GSI of Chinook, coho, and sockeye salmon. The shift to a new technology will require considerable effort and cost, and hence, must be undertaken carefully. The choice of a marker type for application to large-scale management problems must be made in view of several criteria. Among these are greater or equal population resolution than provided by existing markers, ease and cost of genotyping, and suitability of the markers to continue in a well-established tradition of salmon research using genetic tools to provide insights into the breeding biology and genetic population structures of salmon.

Allozyme and microsatellite markers have proved useful in the management of Chinook salmon in several regions stretching from Alaska to California. The development of a coast-wide Chinook salmon database (GAPS) has been the foundation for mixed stock analysis in areas where harvests potentially impact spawning populations in several jurisdictions. This baseline provides helpful insights into the population biology, migration patterns and distributions of Chinook salmon along the coast. At present, a greater number of GSI problems can be addressed by microsatellite baselines than by SNP markers, because of the greater availability of microsatellite markers. Hence, several questions of interest to management presently can be addressed only with microsatellite markers. Microsatellites are usually chosen when both marker types provide similar resolution and are presently the only marker type with proven capability for coast-wide, real-time applications.

Nevertheless, SNPs hold promise for numerous applications. Studies in several taxonomic groups have demonstrated that when SNPs are chosen judiciously, small numbers of SNPs can carry sufficient resolving power for a wide variety of applications, including pedigree analysis in bovids (Werner *et al*. 2003), individual identification in wolves (Seddon *et al*. 2005), cultivar identification for breeding studies in rice (Shirasawa *et al*. 2006), and population genetics in humans (Lui *et al*. 2005). Smith and Seeb (submitted) have pioneered the use of SNP markers in population studies of Pacific salmon. However, it remains to be seen whether SNP markers should replace existing markers for GSI applications. In the long-term, improvements in technology may reduce the price of DNA sequencing so that GSI applications could rely directly on sequence data, rendering genotype assays unnecessary.

SNP markers may improve resolution of some management issues, presently addressed solely with microsatellites. Opportunities for greater resolution may prompt the use of both markers for some applications to maximize resolution. The use of a single marker type will depend upon the resolution provided and cost of analysis of an individual fish. Direct comparisons between microsatellites and SNPs for salmon stock identification applications have been limited to date, as SNP baselines are still under development. SNPs showing large amounts of resolution may be under natural selection and periodic updating of population databases may be important during periods of rapid climate change.

A PSC Expert Panel (Expert Panel PSC 2005) recommended that an evaluation be made of a transition to the use of SNPs for stock identification. Further research is required to determine whether SNPs are capable of outperforming, or meeting the current levels of performance, of microsatellite loci not only for analyzing coast-wide fishery samples, but also for understanding the biology of spawning populations. To help in this decision process, larger SNP databases are required to allow empirical evaluations of resolution. Collaborative projects are underway to collect duplicate tissues for laboratories in the U.S. and Canada. At the same time, established microsatellite population baselines should be maintained and used to aid harvest management.

An important next step is the empirical evaluation of the resolving power for the two markers. This may best be accomplished by focusing closely on one or two species (*e.g.* Chinook and sockeye salmon), for which coast-wide microsatellite baselines (standardized for Chinook) are available, and for which there is a growing SNPs database (largely developed by ADFG). A coast-wide evaluation has not been possible because SNP databases are largely limited to regional population baselines. Importantly, empirical evaluations should include simulations that merge highly informative markers of both classes, as the combination of microsatellite and SNP markers in a single baseline may offer the greatest resolution.

CITATIONS

Allendorf, W.F., Phelps, S.R. 1980. Loss of genetic variation in a hatchery stock of cutthroat trout. *Transactions of the American Fisheries Society* 109: 537–543.

Beacham, T.D., Candy, J.R., Supernault, K.J., Ming, T., Deagle, B., Schultz, A., Tuck, D., Kaukinen, K., Irvine, J.R., Miller, K.M., Withler, R.E. 2001. Evaluation and application of microsatellite and major histocompatibility complex variation for stock identification of coho salmon in British Columbia. *Transactions of the American Fisheries Society* 130: 1116–1155.

Beacham, T.D., Supernault, K.J., Wetklo, M., Deagle, B., Labaree, K., Irvine, J.R., Miller, K.M., Nelson, R.J., Withler, R.E. 2003. The geographic basis for population structure in Fraser River Chinook salmon (Oncorhynchus tshawytscha). *Fishery Bulletin* 101: 229–242.

Beacham, T.D., Lapointe, M., Candy, J.R., McIntosh, B., MacConnachie, C., Tabata, A., Kaukinen, K., Deng, L., Miller, K.M., Withler, R.E. 2004a. Stock identification of Fraser River sockeye salmon using microsatellites and major histocompatibility complex variation. *Transactions of the American Fisheries Society* 133: 1117–1126.

Beacham, T.D., Lapointe, M., Candy, J.R., Miller, K.M., Withler, R.E. 2004b. DNA in action: Rapid application of DNA variation to sockeye salmon fisheries management. *Conservation Genetics* 5: 411–416.

Beacham, T.D., Candy, J.R., McIntosh, B., MacConnachie, C., Tabata, A., Kaukinen, K., Deng, L., Miller, K.M., Withler, R.E., Varnavskaya, N. 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., Candy, J.R., Jonsen, K.L., Supernault, J., Wetklo, M., Deng, L., Miller, K.M., Withler, R.E. 2006a. 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.

Beacham, T.D., Winther, I., Jonsen, K.L., Wetklo, M., Deng, L., Candy, J.R. 2007a. The application of rapid microsatellite-based stock identification to management of a Chinook salmon troll fishery off the Queen Charlotte Islands, British Columbia. *North American Journal of Fisheries Management*. In press.

Beacham, T.D., Wetklo, M., Wallace, C., Olsen, J.B., Flannery, B.G., Wenburg, J.K., Templin, W.D., Antonovich, A., Seeb, L.W. 2007b. The application of microsatellites for stock identification of Yukon River Chinook salmon. *North American Journal of Fisheries Management*. In press.

Bernachetz, L., Duschene, P. 2000. Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles. *Canadian Journal of Fisheries and Aquatic Sciences* 57: 1–12.

Busack, C.A., Currens, K.P. 1995. Genetic risks and hazards in hatchery operations: Fundamental concepts and issues. *American Fisheries Society Symposium* 15: 71–80.

Fischer, S.B., Lerman, L.S. 1983. DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: Correspondence with melting theory. *Proceedings of the National Academy of Science, USA* 80: 1579–1583.

Flannery, B.G., Wenburg, J.K., Gharrett, A.J. 2007. Variation of amplified fragment length polymorphisms (AFLP) in Yukon River chum salmon, *Oncorhynchus keta*: population structure and application to mixed-stock analysis. *Transactions of the American Fisheries Society* In press.

Ford, M.J. 2002. Applications of selective neutrality tests to molecular ecology. *Molecular Ecology* 11: 1245–1262.

Garza, J.C., Williams, E.G. 2001. Detection of reduction in population size using data from microsatellite loci**.** *Molecular Ecology* 10: 305–318.

Grant, W.S. (ed.) 1997. *Genetic effects of straying of non-native hatchery fish into natural populations*: *Proceedings of the workshop June 1-2, 1995.* U.S. Department of Commerce, NOAA Technical Memorandum, NMFS-NWFSC-30.

Kalinowski, S.T. 2004. Genetic polymorphism and mixed stock fisheries analysis. *Canadian Journal of Fisheries and Aquatic Sciences* 61: 1075–1082.

Kwok, P.-Y. (ed.) 2003. *Single Nucleotide Polymorphisms - methods and protocols*. Totowa, New Jersey: Humana Press Inc., 212 p.

LaHood, E.S., Moran, P., Olsen, J., Grant, W.S., Park, L.K. 2002. Microsatellite allele ladders in two species of Pacific salmon: preparation and field-test results. *Molecular Ecology Notes* 2: 187–190.

Liu, N., Chen, L., Wang, S., Oh, C., Zhao, H. 2005. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* 6 (supplement I).

Luikart, G., Allendorf, F.W., Cornuet, J.-M., Sherwin, W.B. 1998. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *Journal of Heredity* 89: 238–247.

Moran, P., Dightman, D.A., Waples, R.S., Park, L.K. 1997. PCR-RFLP analysis reveals substantial population-level variation in the introns of Pacific salmon (*Oncorhynchus spp.*). *Molecular Marine Biology and Biotechnology* 6: 315–327.

Ryman, N., Ståhl, G. 1980. Genetic changes in hatchery stocks of brown trout (*Salmo trutta*). *Canadian Journal of Fisheries and Aquatic Sciences* 37: 82–87.

Ryman, N., Palm, S., André, C., Carvalho, G.R., Dahlgren, T.G., Jorde, P.E., Laikre, L., Larsson, L.C., Palmé, A., Ruzzante, D.E. 2006. Power for detecting genetic divergence: differences between statistical methods and marker loci. *Molecular Ecology* 15: 2031–2045.

Schlötterer, C. 2002. Towards a molecular characterization of adaptation in local populations. *Current Opinion in Genetics and Development* 12: 683–687.

Seeb, L.W., Habicht, C., Templin, W.D., Tarbox, K.E., Davis, R.Z., Brannian, L.K., Seeb, J.E. 2000. Genetic diversity of sockeye salmon of Cook Inlet, Alaska, and its application to management of populations affected by the *Exxon Valdez* oil spill. *Transactions of the American Fisheries Society* 129: 1223–1249.

Seeb, L.W., Crane, P.A., Kondzela, C.M., Wilmot, R., Urawa, S.N. 2004. Migration of Pacific Rim chum salmon on the high seas: insights from genetic data. *Environmental Biology of Fishes* 69: 21–36.

Seeb, L.W., Antonovich, A., Banks, M.A., Beacham, T.D., Bellinger, M.R., Campbell, M., Garza, J.C., Guthrie III, C.M., Moran, P., Narum, S.R., Stephenson, J.J., Supernault, K.J., Teel, D.J., Templin, W.D., Wenburg, J.K., Young, S.F., Smith, C.T. Submitted. Development of a Standardized DNA Database for Chinook Salmon. *Fisheries*.

Seddon, J.M., Parker, H.G., Ostrander, E.A., Ellegren, H. 2005. SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology* 14: 503–511.

Shirasawa, K., Shiokai, S., Yamaguchi, M., Kihsitani, S., Nishio, T. 2006, Dot-blot-SNP analysis for practical plant breeding and cultivar identification in rice. *Theoretical Applications in Genetics* 113: 147–155.

Small, M.P., Withler, R.E., Beacham, T.D. 1998. Population structure and stock identification of British Columbia coho salmon, *Oncorhynchus kisutch*, based on microsatellite DNA variation. *Fishery Bulletin* 96: 843–858.

Smith, C.T., Seeb, L.W. Submitted. Number of alleles as a predictor of the relative assignment power of SNP and STR baselines for chum salmon. *Transactions of the American Fisheries Society*.

Smith, C.T., Koop, B.F., Nelson, R.J. 1998. Isolation and characterization of coho salmon (Oncorhynchus kisutch) microsatellites and their use in other salmonids. *Molecular Ecology* 7: 1614–1616.

Smith, C.T., Antonovich, A., Templin, W.D., Narum, S.R., Elfstrom, C.M., Seeb, L.W. In press. Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon; a comparison of SNPs to STRs and allozymes. *Transactions of the*

American Fisheries Society.
Smith, C.T., Elfstrom, C.M., Seeb, J.E., Seeb, L.W. 2005a. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology* 14: 4193–4203.
Smith, C.T., Templin, W.D., Seeb, J.E., Seeb, L.W. 2005b. Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of U.S. and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management* 25: 944–953.
Smith, C.T., Baker, J., Park, L., Seeb, L.W., Elfstrom, C., Abe, S., Seeb, J.E. 2005c. Characterization of 13 single nucleotide polymorphism markers for chum salmon. *Molecular Ecology Notes* 5: 259–262.
Smith, C.T., Seeb, J.E., Schwenke, P., Seeb, L.W. 2005d. Use of the 5'-nuclease reaction for SNP genotyping in Chinook salmon. *Transactions of the American Fisheries Society* 134: 207–217.
Smith, C.T., Park, L., Van Doornik, D., Seeb, L.W., Seeb, J.E. 2006. Characterization of 19 single nucleotide polymorphism markers for coho salmon. *Molecular Ecology Notes* 6: 715–720.
Smith, C.T., Antonovich, A., Templin, W.D., Elfstrom, C.M., Narum, S.R., Seeb, L.W. 2007. Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon; a comparison of SNPs to STRs and allozymes. *Transactions of the American Fisheries Society*. In Press
Spidle, A.P., Quinn, T.P., Bentzen, P. 2000. GenBank accession number AF272822. Oncorhynchus kisutch microsatellite Oki23 sequence. Available: http://www.ncbi.nlm.nih.gov. (April 2006).
Teel, D.J., Milner, G.B., Winans, G.A., Grant, W.S. 2001. Genetic population structure and origin of life history types in Chinook salmon in British Columbia, Canada. *Transactions of the American Fisheries Society* 129: 194–209.
Teel, D.J., Van Doornik, D.M., Kuligowski, D.R., Grant, W.S. 2003. Genetic analysis of juvenile coho salmon (*Oncorhynchus kisutch*) off Oregon and Washington reveals few Columbia River wild fish. *Fishery Bulletin* 101: 640–652.
Van Doornik, D.M., Teel, D.J., Kuligowski, D.R., Morgan, C.A., Casillas, E. 2007. Genetic analysis provide insight into early ocean stock distribution and survival of survival of juvenile coho salmon off the coasts of Washington and Oregon. *North American Journal of Fisheries Management* 27: 220–237.
Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23: 4407–4414.
Waples, R.S. 1990. Conservation genetics of Pacific salmon. III. Estimating effective population size. *Journal of Heredity* 81: 277–289.
Werner, F.A.O., Durstewitz, G., Habermann, F.A., Thaller, G., Kramer, W., Kollers, S., Buitkamp, J., Georges, M., Brem, G., Mosner, J., Fries, R. 2004. Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Animal Genetics* 35: 44–49.
Winther, I., Beacham, T.D. 2006. The application of Chinook salmon stock composition data to management of the Queen Charlotte Islands troll fishery, 2002 to 2005. *Canadian Technical Report of Fisheries and Aquatic Sciences* 2665: 88 p.

APPENDIX B. STATUS OF TISSUE COLLECTIONS AND MOLECULAR MARKERS FOR COHO AND SOCKEYE SALMON

APPENDIX B. Status of tissue collections and molecular markers for coho and sockeye salmon

Chris Habicht
Gene Conservation Laboratory, Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, AK 99518

Terry Beacham
Pacific Biological Station, Department of Fisheries and Oceans, Nanaimo, BC, V9R 5K6

Christian Smith
Abernathy Fish Technology Center, US Fish and Wildlife Service, 1440 Abernathy Creek Road, Longview, WA 98632

Don Van Doornik, Eric Iwamoto
Conservation Biology Division, NOAA Fisheries, 2725 Montlake Boulevard East, Seattle, WA 98112

INTRODUCTION

One step in constructing databases for salmon populations is the development of regional baselines, usually by agencies with regional management mandates. The developments of coast-wide baselines often arise from these regional baselines following a standardization process. Coast-wide allozyme baselines were developed and maintained through collaborations and workshops to describe and standardize allele nomenclatures. Development of a coast-wide microsatellite baseline for Chinook salmon was a continuation of this process, but represented a considerable advance in standardization protocols and data access by making much of the data accessible over the internet.

The term 'standardization' as used here may describe different levels of cooperation among laboratories. For example, several laboratories may share a set of markers and allele designations, but no common set of population baseline data (*e.g.* present state of the steelhead baseline), or some laboratories may share markers, allelic nomenclature and baseline data (*e.g.* present state for coho salmon). While these two scenarios represent steps toward standardized coast-wide databases, neither is as valuable to the Pacific Salmon Commission as the GAPS Chinook baseline standardization, which ties together the activities of at least 12 laboratories.

The goal of this section is to summarize existing regional and coast-wide datasets with a focus on coho and sockeye salmon. The following tables of data are snapshots of a growing set of regional and coast-wide databases for SNP and microsatellite markers.

MICROSATELLITE BASELINE DATA

Coho salmon

Populations of coho salmon have been sampled by DFO in several areas and examined for 13 microsatellite loci and two MHC exons. These samples are concentrated in British Columbia with representative samples from Southeast Alaska and from Washington State (Tables B1, B2). A large number of samples have been examined for variability at 11 microsatellite loci by NOAA Fisheries, Seattle from populations extending from southern British Columbia to northern California (Tables B3, B4). Presently, 61 microsatellite primers developed for coho or other species of salmon have been used to screen for variability in coho salmon (Table B5).

Table B1. *Coho salmon.* DFO: Regions and populations within regions included in the survey of variation at 13 microsatellite loci and two MHC exons in coho salmon. Number in parentheses after the name refers to the location shown in Figure 1 in Beacham et al. (2001)

Region	Subregion	Number of populations	Populations
Southeast Alaska		9	Berners (1), Gastineau Hatchery (2), Hidden Falls (3), Reflection Lake (4), Indian Creek (5), Margaret Creek (6), Karta (7), Whitman Lake (8), Hugh Smith (9)
Queen Charlotte Islands	North coast	3	Sangan River (10), Awun River (11), Yakoun River (12)
	East coast	3	Deena (13), Copper River (14), Pallant Creek (15)
	West coast	1	Tasu (16)
Nass River		3	Meziadin (17), Zolzap (18), Tseax (19)
North coast		1	Lachmach (20)
Upper Skeena River	Upper drainage	3	Kluatantan (21), Sustut River (22), Motase (23)
	Babine River	3	Babine Fence (24), Boucher (25), Upper Babine (26)
	Bulkley River	4	Toboggan Creek (27), Bulkley River (28), Morice River (29), Owen (30)
Lower Skeena River	Mid drainage	3	Kispiox (31), Kitwanga (32), Singlehurst (33)
	Lower drainage	14	Hadenschild (34), Cedar (35), Clear (36), Deep (37), Kitsumkalum (38), Zymagotitz (39), Sockeye (40), Schulbuckhand (41), Clearwater (42), Coldwater (43), Exchamsiks (44), Kasiks (45), Green (46), Ecstall (47)
Central Coast		11	Kitimat (48), Hartley Bay (49), Kitasoo (50), McLaughlin Bay (51), Atnarko (52), Salloomt (53), Thorsen (54), Sheemahant (55), Docee (56), Devereux (57), Klinaklini (58)
Northern Vancouver Island		8	Nahwitti (59), Wanokana (60), Stephens (61), Quatse (62), Waukwass (63), Cluxewe (64), Glen Lyon (65), Nimpkish (66)
Eastern Vancouver Island		8	Quinsam (67), Black Creek (68), Puntledge (69), Big Qualicum (70), Nanaimo (71), Chemainus (72), Cowichan (73), Goldstream (74)
Western Vancouver Island		13	Conuma (75), Cypre (76), Tranquil (77), Kennedy (78), Kootowis (79), Robertson Creek (80), Sarita (81), Pachena (82), Nitinat (83), San Juan (84), Kirby Creek (85), Sooke (86), Craigflower (87)
Southern mainland		6	Homathko (88), Lang Creek (89), Sliammon (90), Squamish (91), Seymour (92), Capilano River (93)
Fraser River	Lower drainage	10	Pitt (94), Alouette (95), Stave (96), Inch Creek (97), Norrish (98), Nicomen (99), Chehalis (100), Chilliwack (101), Kanaka Creek (102), Salmon River (103)

	Upper drainage	2	Bridge River (104), McKinley Creek (105)
Thompson River	Lower drainage	3	Spius Creek (106), Coldwater (107), Deadman (108)
	North Thompson	5	Louis Creek (109), Dunn Creek (110), Lemieux Creek (111), Mann (112), Lion (113)
	South Thompson	7	Momich (114), Eagle (115), Salmon (116), Danforth (117), Duteau (118), Bessette (119), Lang Channel (120)
Puget Sound		6	Nooksack (121) Marblemount (122), Wallace (123), Grizzly (124), Minter (125), Nisqually (126)
Hood Canal		1	Dewatto (127)
Juan de Fuca		2	Dungeness (128), Elwha (129)
Coastal		6	Quillayute (130), Clearwater (131), Shale (132), Queets (133), Bingham (134), Willapa (135)
Columbia River		3	Cowlitz (136), Lewis (137), Clackamas (138)

Table B2. *Coho salmon.* Regions, number of collections within regions, and number of individuals included in the survey of variation at 13 microsatellite loci and two MHC exons in coho salmon (T. Beacham, DFO)

Region	Number of collections	Number of individuals
Transboundary	7	700
SE Alaska	9	1450
QCI	20	1400
Nass	3	750
Skeena	29	4500
Central Coast	40	5850
South Coast	28	3650
ECVI	22	6350
WCVI	13	4300
Fraser	47	13,300
Washington	15	1400
Columbia	9	800
Oregon	10	850
California	4	150

Table B3. *Coho salmon.* NOAA Fisheries, Seattle: Population samples analyzed for variation at 11 core microsatellite loci listed in Table B5). [From Van Doornik et al. (2007)]

Region Subregion Location	Sample size	Life stage
British Columbia		
West coast Vancouver Island		
1 Tranquil Creek Hatchery	81	Adult
2 Upper Kennedy River Hatchery	72	Adult
3 Nitinat River Hatchery	95	Adult
East coast Vancouver Island		
4 Nanaimo River Hatchery	96	Adult
5 Cowichan Fish Hatchery	89	Adult
6 Goldstream Salmon Hatchery	96	Adult
Southern BC coast		
7 Homathko River	73	Adult
8 Tenderfoot Creek Hatchery	91	Adult
9 Capilano Salmon Hatchery	79	Adult
Lower Fraser River		
10 Inch Creek Hatchery	78	Adult
11 Chehalis River Hatchery	87	Adult
12 Chilliwack Hatchery	82	Adult
Mid-Fraser–Thompson River		
13 Dunn Creek	76	Adult
14 Bridge Creek	90	Adult
15 Bessette Creek	79	Adult
Puget Sound		
Puget Sound without Hood Canal		
16 Nooksack Hatchery	95	Parr
17 Ennis Creek, Samish River	140	Adult
18 Skagit Hatchery	87	Parr
19 Fortson Creek	41	Adult
20 Grizzly Creek, Snoqualmie River	105	Adult
21 Soos Creek Hatchery	450	Adult, parr
22 Minter Creek Hatchery	40	Adult
Hood Canal		
23 Quilcene Hatchery	141	Adult
24 Rockybrook Creek, Dosewallips River	32	Adult
25 Big Beef Creek	134	Adult, smolt
26 Hatchery Creek, Duckabush River	78	Adult

27 John Creek, Hamma Hamma River	86	Adult
28 Dewatto River	115	Adult
29 George Adams Hatchery	91	Adult
30 Kirkland and Fir Creek, Skokomish River	94	Adult
Washington coast		
Strait of Juan de Fuca		
31 Snow Creek	137	Adult
32 Dungeness Hatchery	47	Parr
33 Elwha Hatchery	186	Adult
North Washington coast		
34 Hoko River	76	Adult
35 Makah Hatchery	143	Adult
36 Sol Duc Hatchery (summer run)	96	Parr
36 Sol Duc Hatchery (fall run)	94	Parr
36 Sol Duc River (summer run)	95	Parr
37 Clearwater River	117	Adult, smolt
38 Queets River	156	Adult, parr
39 Quinault Hatchery	139	Adult
South Washington coast		
40 Humptulips Hatchery (early run)	47	Parr
41 Bingham Creek Hatchery, Chehalis River	66	Parr
42 Hope Creek, Chehalis River	44	Parr
43 Nemah Hatchery	94	Parr
44 Naselle Hatchery	94	Parr
Columbia River		
45 Elochoman Hatchery (early run)	42	Parr
45 Elochoman Hatchery (late run)	46	Parr
46 Cowlitz Hatchery	137	Parr
47 Fallert Creek (Kalama) Hatchery (early run)	92	Parr
47 Kalama Falls Hatchery (late run)	83	Parr
48 Lewis Hatchery (early run)	46	Parr
48 Lewis Hatchery (late run)	48	Parr
49 Big Creek Hatchery	88	Parr
50 Clackamas River (early run)	54	Adult
50 Clackamas River (late run)	31	Adult
51 Eagle Creek Hatchery	96	Adult
52 Sandy Hatchery	95	Parr
53 Bonneville Hatchery	94	Parr
Oregon coast region		
North-central Oregon coast		
54 Nehalem Hatchery	92	Parr
55 Trask Hatchery	94	Parr
56 Devil's Lake	60	Adult
57 Siletz River	69	Adult
58 Yaquina River	66	Adult

59 Beaver Creek	64	Adult
60 Alsea River	62	Adult
61 Siuslaw River	150	Adult
62 Coos River	76	Parr
63 Bethel Creek, New River	30	Parr
Oregon lakes complex		
64 Sutton Creek	48	Adult
65 Mercer Lake	28	Adult
66 Siltcoos Lake	53	Adult
67 Tahkenitch Lake	34	Adult
68 Ten Mile Lake	75	Adult, parr
Umpqua River		
69 Mainstem Umpqua River	53	Adult
70 Smith River, Umpqua River	128	Adult, parr
71 Elk Creek, Umpqua River	30	Adult
72 Calapooya River, Umpqua River	34	Adult
73 Rock Creek, North Umpqua River	55	Parr
74 South Fork, Umpqua River	67	Adult
South Oregon–north California coasts		
75 Elk River	23	Parr
76 Cole Rivers Hatchery (Rogue stock)	34	Parr
77 Irongate Hatchery	106	Parr
78 Trinity River Hatchery	102	Parr

Table B4 *Coho salmon.* NOAA Fisheries, Seattle: Microsatellie loci, annealing temperatures and primer references used to evaluate stock composition. [from Van Doornik et al. (2007)]

Locus	Annealing temperature	Reference
Ocl8	60	Condrey and Bentzen (1998)
Oki1	58	Smith et al. (1998)
Oki10	60	Smith et al. (1998)
Oki23	58	Spidle et al. (2000)
One13	58	Scribner et al. (1996)
Ots3	47	Banks et al. (1999)
Ots103	54	Small et al. (1998)
Ots213	58	Greig et al. (2003)
Ots505 NWFSC	54	Naish and Park (2002)
OtsG422	58	Williamson et al. (2002)
P53	58	de Fromentel et al. (1992)

Table B5. *Coho salmon.* Status of screening for microsatellites among laboratories as of July 2007 (compiled by D. Van Doornik, NOAA Fisheries)

Locus	NMFS Manchester	NMFS Santa Cruz	CDFO/WDFW collaboration	USFWS Abernathy	USFWS Alaska	OSU	BML	Allele ladder candidates
Ocl8	X	X	X	X				1
Oki1	X	X	X	X	X		X	1
Oki10	X		X	X				2
Oki23	X			X		X		3
One13	X	X		X		X	X	2
Ots103	X	X	X	X		X	X	1
Ots213	X		X	X		X		2
Ots3	X			X		s	X	3
OtsB3	X			X				3
OtsG422	X	X		X				2
P53	X		X	X		X	X	1
iso-Ots2	S						X	
Oki11	S				X			
Oki13	S	X						
Oki2	S							
Oki3	S				X			
Ots101	S		X					3
Ots105	S	X			X			3
Ots2	S					X	X	
Ots208	S							
Ots212	S					s		
OtsG249	S							
OtsG253b	S		X					3
OtsG3	S	X						
OtsG68	S	X						
OtsG78b	S	X						
OtsG83b	S	X						
Ogo1a								
Ogo2			X					3
Oke2					X			
Oke3					X			
Oke4					X			
Oki100			X					3
Oki101			X					3
Oki16						X		
Omm1121								
Omm1128								
Omy1011			X					3
Omy116		X						
Omy325			X					3
Omy77						s		

One111			X					3
One11b		X						
One13M			X					3
One2								
One3					X			
Ots1						s		
Ots10						s		
Ots108		X						
Ots1b		X						
Ots206						s		
Ots208b						s		
Ots209						s		
Ots215						X		
Ots2M			X					3
Ots3.1					X			
Ots3M			X					3
Ots9						s		
Ssa14		X						
Ssa407			X					3
Ssa85		X						
Total in use	11	17	18	11	9	8	7	

X = locus is in use

s = locus has been screened and is being or has been evaluated for possible use

Sockeye salmon

Several regional databases for microsatellite markers in sockeye salmon have been used by DFO (Table B6) and NOAA (TableB7). Most surveys of microsatellite loci have been of populations in British Columbia (Table B8), and only of a few populations of conservation concern in Washington (Table B7).

Table B6 *Sockeye salmon.* DFO: Summary of microsatellite markers available and number of observed alleles recorded by the DFO laboratory (T. Beacham)

Microsatellite locus	Number of alleles
Oki1a	8
Oki1b	10
Ots107	15
Omy77	20
Ots2	26
Ots3	26
Oki16	26
Ots108	29
Ots103	30
One8	32
Ots100	33
Oki6	37
Oki29	39
Oki10	83
DAB-β1	15

Table B7. *Sockeye salmon.* NOAA Fisheries, Seattle: Data from Redfish Lake and the Wenatchee and Okanagan rivers are available for the following microsatellite loci (E. Iwamoto, NOAA Fisheries, Seattle)

Locus
Oke2
One110
Omm1085
One18
Ots10M
Ots100
Ssa85
Ots519
One13
Omm 1159
Omy77
Ots103
Ots3
One21
Omm1068
Oki29

Table B8. *Sockeye salmon.* DFO: Summary of the number of sampling sites or populations within geographic regions. A complete listing of the populations is outlined by Beacham et al. (2005) in their Appendix Table 1. Range of annual and population samples sizes within regions is in parentheses. Fourteen microsatellite loci and an MHC locus were surveyed as outlined by Beacham et al. (2005)

Region	Number of populations	Mean annual sample size	Mean population sample size
Columbia River	2	71 (15, 194)	285 (68,502)
Washington	3	114 (50, 201)	114 (50, 201)
Fraser River	53	94 (5, 400)	270 (15, 858)
West coast Vancouver Island	15	90 (19, 197)	132 (19, 279)
Nimpkish River	3	108 (42, 290)	288 (203, 367)
Southern BC	6	114 (12, 219)	171 (18, 325)
Central BC	16	79 (27, 223)	97 (27, 223)
Owikeno Lake	10	77 (7, 114)	224 (86, 398)
Long Lake	3	99 (39, 205)	297 (139, 490)
Queen Charlotte Islands	5	71 (41, 99)	114 (41, 190)
Nass River	11	96 (24, 264)	313 (40, 797)
Skeena River	14	78 (33, 200)	151 (33, 287)
Babine Lake	11	95 (54, 200)	208 (78, 499)
Unuk River	1	50 (50,50)	50 (50,50)
Stikine River	17	83 (6, 405)	152 (26, 474)
Taku River	10	57 (12, 100)	86 (12, 199)
Alsek River	15	83 (10, 238)	144 (10, 592)
Southeast Alaska	20	151 (45, 343)	197 (45, 300)
Kodiak Island	15	73 (15, 112)	73 (15, 112)
Bristol Bay	14	76 (47, 101)	97 (50, 153)
Alaska Peninsula	2	88 (75, 100)	88 (75, 100)
Chukotka	8	25 (20, 30)	25 (20, 30)
Olutorsky Bay	5	75 (48, 180)	105 (48, 180)
Navarinsky Region	1	100 (100, 100)	100 (100, 100)
Karaginsky Bay	1	98 (98, 98)	98 (98, 98)
Kamchatka River	16	58 (15, 120)	72 (15, 190)
Kronotzky Bay	1	44 (44, 44)	44 (44, 44)
Southeast Kamchatka	3	48 (35, 71)	48 (35, 71)
Kurilskoye Lake	12	58 (35, 103)	78 (50, 121)
Southwest Kamchatka	1	52 (52, 52)	52 (52, 52)
Bolshaya River	4	56 (25, 97)	84 (25, 147)
Tigil River	1	101 (101, 101)	101 (101, 101)
Palana River	1	49 (49, 49)	49 (49, 49)
Hokkaido Island	1	75 (75, 75)	75 (75, 75)

SNP BASELINE DATA

Most SNP databases encompass only regional sets of populations. Presently, 51 genotypic assays are available for Chinook salmon, 19 for coho salmon, 77 for chum salmon, 44 for sockeye salmon and none for pink salmon (Tables B9, B10a). The numbers of SNP assays and the numbers of samples examined is growing rapidly. About 35,000 sockeye salmon have been examined for SNP variability (Table B11a) in samples extending from Russia to Washington-Idaho, but with a concentration in Alaska around Bristol Bay and the Alaska Peninsula, where this species is most abundant (Table B11a). About 42 SNP assays have been developed for coho salmon (Table B10b), but only about 400 fish have been examined for variability in samples extending from Russia to Washington (Table B11b). SNP assays have also been developed for chum salmon (n = 77; Tables B9 and B10c) and for Chinook salmon (n = 51; Tables B9 and B10d). About 12,000 chum salmon have been examined for variability in samples extending from Korea to Washington (Table B11c), and nearly 25,000 Chinook salmon have been examined in samples from Russia to California (Table B11d). Several thousand Chinook salmon from Southeast Alaska and the Yukon-Kuskokwim rivers have been examined to support transboundary management.

Table B9. Number of SNP genotyping assays available for each species of Pacific salmon (compiled by C. Smith, USFWS).

Species	Number of available genotyping assays
Chinook salmon	51[1,2,3,4]
Coho salmon	19[5]
Chum salmon	77[1,6,7,8]
Sockeye salmon	44[1,9]
Pink salmon	0

1) Smith *et al*. (2005a), 2) Smith *et al*. (2005d), 3) Smith *et al*. (in press), 4) Narum *et al*. (in press), 5) Smith *et al*. (2006), 6) Elfstrom *et al*. (in press), 7) Smith *et al*. (2005c), 8) Garvin and Gharrett (in press), 9) Elfstrom *et al*. (2006).

Table B10. Single Nucleotide Polymorphism markers assayed for a) sockeye salmon, b) coho salmon, c) chum salmon, and d) Chinook salmon. Nuclear markers are diploid and mtDNA are haploid (C. Habicht, ADFG).

a. Sockeye salmon

Published name	Ploidy	Reference*
One_ACBP-79	D	1
One_ALDOB-135	D	1
One_CO1	H	1
One_ctgf-301	D	1
One_Cytb_17	H	1
One_Cytb_26	H	1
One_E2-65	D	2
One_GHII-2165	D	1
One_GPDH-201	D	2
One_GPDH2-187	D	2
One_GPH-414	D	1
One_hsc71-220	D	1
One_HGFA-49	D	2
One_HpaI-71	D	1
One_HpaI-99	D	1
One_IL8r-362	D	3
One_KPNA-422	D	1
One_LEI-87	D	1
One_MARCKS-241	D	3
One_MHC2_190	D	1
One_MHC2_251	D	1
One_Ots213-181	D	1
One_p53-534	D	1
One_ins-107	D	2
One_Prl2	D	1
One_RAG1-103	D	1
One_RAG3-93	D	1
One_RFC2-102	D	2
One_RFC2-285	D	2
One_RH2op-395	D	1
One_serpin-75	D	2
One_STC-410	D	1
One_STR07	D	1
One_Tf_ex11-750	D	1
One_Tf_in3-182	D	1
One_U301-92	D	1
One_U401-224	D	3
One_U404-229	D	3
One_U502-167	D	3

One_U503-170	D	3
One_U504-141	D	3
One_U508-533	D	3
One_VIM-569	D	1
One_ZNF-61	D	3
One_Zp3b-49	D	2

* 1) Elfstrom et al. (2006), 2) Smith et al. (2005a), 3) Alaska Department of Fish and Game (unpublished)

b. Coho salmon

Published name	Ploidy	Reference**
Oki_arf-115	D	1
Oki_BAMBI-128	D	2
Oki_BAMBI-172	D	2
Oki_CR-209	H	1
Oki_CR-296	H	1
Oki_E2-84	D	1
Oki_eif4ebp2-148	D	2
Oki_eif4ebp2-58	D	1
Oki_GnRH-151	D	1
Oki_GPDH-146	D	1
Oki_GPDH-187	D	1
Oki_HGFA-311	D	1
Oki_IGF-I.1-163	D	1
Oki_ins-167	D	1
Oki_ins-323	D	1
Oki_LWSop-554	D	1
Oki_RACP-176	D	1
Oki_SClkF2R2-120	D	1
Oki_serpin-130	D	1
Oki_serpin-328	D	1
Oki_SWS1op-38	D	1
Oki_u6-258	D	1

**1) Smith et al. (2006), 2) Alaska Department of Fish and Game (unpublished)

c. Chum salmon

Published name	Ploidy	Reference***
Oke_PPA2-635	D	1
Oke_AhR1-278	D	1
Oke_AhR1-78	D	1
Oke_arf-319	D	2
Oke_U401-143	D	1
Oke_U401-220	D	1
Oke_CKS-389	D	3
Oke_copa-211	D	2
Oke_Cr30	H	3
Oke_Cr386	H	3

Oke_ctgf-105	D	1
Oke_DM20-548	D	3
Oke_eif4ebp2-64	D	2
Oke_FARSLA-242	D	1
Oke_GHII-2943	D	1
Oke_GHII-3129	D	1
Oke_GnRH-373	D	3
Oke_GnRH-527	D	3
Oke_GPDH-191	D	2
Oke_GPH-105	D	1
Oke_GPH-78	D	1
Oke_hnRNPL-239	D	1
Oke_HP-182	D	1
Oke_HSP90BA-299	D	1
Oke_hsc71-199	D	2
Oke_il-1racp-67	D	2
Oke_IL8r-272	D	3
Oke_IL8r-406	D	3
Oke_KPNA2-87	D	1
Oke_MAPK1-135	D	1
Oke_MARCKS-362	D	1
Oke_Moesin-160	D	2
Oke_ND3-69	H	3
Oke_ras1-249	D	1
Oke_RFC2-618	D	2
Oke_RH1op-245	D	2
Oke_serpin-140	D	2
Oke_TCP1-78	D	1
Oke_Tf-278	D	1
Oke_Tsha1-196	D	2
Oke_u1-519	D	3
Oke_u202-131	D	2
Oke_u212-87	D	2
Oke_u216-222	D	2
Oke_u217-172	D	2
Oke_u200-385	D	2
Oke_U302-195	D	1
Oke_U502-241	D	1
Oke_U503-272	D	1
Oke_U504-228	D	1
Oke_U505-112	D	1
Oke_U506-110	D	1
Oke_U507-286	D	1
Oke_U507-87	D	1
Oke_U509-219	D	1
Oke_U510-204	D	1
Oke_U511-271	D	1
Oke_U514-150	D	1
Oke_U305-130	D	1

Oke_U305-307	D	1

***1) Elfstrom et al. (in press), 2) Smith et al. (2005a), 3) Smith et al. (2005b)

d. Chinook salmon

Published Name	Ploidy	Reference****
GTH2B-550	D	1
NOD1	D	1
Ots_E2-275	D	2
Ots_arf-188	D	2
Ots_AsnRS-60	D	2
Ots_C3N3	H	2
Ots_E9BAC	D	1
Ots_ETIF1A	D	1
Ots_FARSLA-220	D	3
Ots_FGF6A	D	1
Ots_FGF6B	D	1
Ots_GH2	D	2
Ots_GPDH-338	D	2
Ots_GPH-318	D	3
Ots_GST-207	D	3
Ots_GST-375	D	3
Ots_HGFA-446	D	2
Ots_hnRNPL-533	D	3
Ots_HSP90B-100	D	3
Ots_HSP90B-385	D	3
Ots_IGF-I.1-76	D	2
Ots_Ikaros-250	D	2
Ots_il-1racp-166	D	2
Ots_LEI-292	D	3
Ots_MetA	D	1
Ots_MHC1	D	2
Ots_MHC2	D	2
Ots_ZNF330-181	D	2
Ots_LWSop-638	D	2
Ots_SWS1op-182	D	2
Ots_P450	D	2
Ots_P53	D	2
Ots_Prl2	D	2
Ots_ins-115	D	2
Ots_PSMB1-197	D	3
Ots_RFC2-558	D	2
Ots_SClkF2R2-135	D	2
Ots_SERPC1-209	D	3
Ots_SL	D	2
Ots_TAPBP	D	1
Ots_Tnsf	D	2

****1) GAPS (2006), 2) Smith et al. (2005a), 3) Smith et al. (in press)

Table B11. Number of a) sockeye salmon, b) coho salmon, c) chum salmon, and d) Chinook salmon from baseline collections throughout the Pacific Rim that have been screened for all Single Nucleotide Polymorphism markers detailed Table B10. Multilocus genotypes are archived in the Alaska Department of Fish and Game database (C. Habicht, ADFG).

a. Sockeye salmon

Region	Number of samples	Number of individuals
Washington/Idaho	2	193
British Columbia	41	3,347
Southeast Alaska	36	3,244
North Gulf Coast	7	554
Southcentral Alaska	78	8,035
Kodiak and AK Peninsula	74	6,985
Bristol Bay	98	9,770
Arctic-Yukon-Kuskokwim	16	1,046
Russia	40	2,211
Total	393	35,385

b. Coho salmon

Region	Number of samples	Number of individuals
Washington/Idaho	1	96
Southeast Alaska	1	48
Southcentral Alaska	1	94
Bristol Bay	1	54
Arctic-Yukon-Kuskokwim	1	48
Russia	1	38
Total	6	378

c. Chum salmon

Region	Number of samples	Number of individuals
Washington/Idaho	8	281
British Columbia	2	96
Southeast Alaska	11	887
Southcentral Alaska	7	568
Kodiak and AK Peninsula	17	1,307
Bristol Bay	8	636
Arctic-Yukon-Kuskokwim	59	5,606
Russia	13	745
Japan	19	1,532

Korea	2	191
Total	146	11,849

d. Chinook salmon

Region	Number of samples	Number of individuals
California	9	366
Oregon	3	282
Washington/Idaho	11	976
British Columbia	61	5,522
Southeast Alaska	50	3,965
North Gulf Coast	32	1,833
Southcentral Alaska	23	2,190
Kodiak and AK Peninsula	14	864
Bristol Bay	9	480
Arctic-Yukon-Kuskokwim	119	7,837
Russia	8	411
Total	339	24,726

CITATIONS

Beacham, T.D., Candy, J.R., McIntosh, B., MacConnachie, C., Tabata, A., Kaukinen, K., Deng, L., Miller, K.M., Withler, R.E., Varnavskaya, N. 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.

Elfstrom, C.M., Smith, C.T., Seeb, J.E. 2006. Thirty-two single nucleotide polymorphism markers for high-throughput genotyping of sockeye salmon. *Molecular Ecology Notes* 6: 1255–1259.

Elfstrom, C.M., Smith, C.T., Seeb, L.W. In press. Thirty-eight single nucleotide polymorphism markers for high-throughput genotyping of chum salmon. *Molecular Ecology Notes*.

Garvin, M.R., Gharrett, A.J. In press DEco-TILLING: an inexpensive method for single nucleotide polymorphism discovery that reduces ascertainment bias. *Molecular Ecology Notes*.

Narum, S.R., Banks, M., Beacham, T., Bellinger, R., Campbell, M., DeKoning, J., Elz, A., Guthrie, C., Kozfkay, C., Miller, K., Moran, P., Phillips, R., Seeb, L., Smith, C., Warheit, K., Young, S., Garza, J.C. In press. Differentiating populations at broad and fine geographic scales with microsatellites and SNPs. *Molecular Ecology*.

Smith, C.T., Elfstrom, C.M., Seeb, J.E., Seeb, L.W. 2005a. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology* 14: 4193–4203.

Smith, C.T., Templin, W.D., Seeb, J.E., Seeb, L.W. 2005b. Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of U.S. and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management* 25: 944–953.

Smith, C.T., Baker, J., Park, L., Seeb, L.W., Elfstrom, C., Abe, S., Seeb, J.E. 2005c. Characterization of 13 single nucleotide polymorphism markers for chum salmon. *Molecular Ecology Notes* 5: 259–262.

Smith, C.T., Seeb, J.E., Schwenke, P., Seeb, L.W. 2005d. Use of the 5'-nuclease reaction for SNP genotyping in Chinook salmon. *Transactions of the American Fisheries Society* 134: 207–217.

Smith, C.T., Park, L., Van Doornik, D., Seeb, L.W., Seeb, J.E. 2006. Characterization of 19 single nucleotide polymorphism markers for coho salmon. *Molecular Ecology Notes* 6: 715–720.

Smith, C.T., Antonovich, A., Templin, W.D., Elfstrom, C.M., Narum, S.R., Seeb, L.W. In press. Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon; a comparison of SNPs to STRs and allozymes. *Transactions of the American Fisheries Society*.

Van Doornik, D.M., Teel, D.J., Kuligowski, D.R., Morgan, C.A., Casillas, E. 2007. Genetic analysis provide insight into early ocean stock distribution and survival of survival of juvenile coho salmon off the coasts of Washington and Oregon. *North American Journal of Fisheries Management* 27: 220–237.

APPENDIX C. COAST-WIDE INTEGRATION OF GSI DATA COLLECTION, INTERPRETATION, AND USE IN MIXED STOCK ANALYSES

APPENDIX C. Coast-wide integration of GSI data collection, interpretation, and use in mixed stock analyses

W. Stewart Grant
Department of Biological Sciences, University of Alaska Anchorage, Anchorage, AK 99506

Shawn Narum
Columbia River Inter-Tribal Fish Commission, 729 NE Oregon, Suite 200, Portland, OR 97206

Terry Beacham
Pacific Biological Station, Department of Fisheries and Oceans, Nanaimo, BC, V9R 5K6

Lisa Seeb
Gene Conservation Laboratory, Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, AK 99518

Steve Latham
Pacific Salmon Commission, 600 - 1155 Robson Street, Vancouver, BC, V6E 1B5

Paul Moran
Conservation Biology Division, NOAA Fisheries, 2725 Montlake Boulevard East, Seattle, WA 98112

INTRODUCTION

The value of genetic stock identification is greatly enhanced by ensuring that individual datasets can be merged into a larger coast-wide dataset. Unified datasets are especially important for Pacific salmon which often make migrations of several thousand kilometers and which can be harvested in fisheries far removed from spawning areas. Several studies have provided insights into high seas abundance patterns of adults (Seeb *et al*. 2004) and juveniles (Teel *et al*. 2003; Van Doornik *et al*. 2007) in areas far removed from spawning streams and rivers. These data should be accessible in a timely manner to management agencies responsible for maintaining sustainable harvests of salmon. Previous efforts to integrate databases for Chinook salmon (GAPS) have proved successful and have provided insights into the biology of Chinook populations that were not apparent with the separate analyses of individual datasets. The need for greater integration of existing data for other species of Pacific salmon is recognized by the Pacific Salmon Commission and by the Tribal, State and Provincial agencies responsible for harvest management in the Northeastern Pacific.

COMPONENTS OF DATA SHARING

Some of the existing regional or agency datasets cannot be merged to provide a broader picture for a particular species because of differences in sampling or laboratory protocols. The use of genetic data from several laboratories requires attention to several layers of detail to be able to merge datasets to provide a broad geographic perspective on genetic population structure and to use in mixed stock analyses (Moran *et al*. 2006).

Common set of loci must be examined among laboratories for each class of molecular marker

The first criterion requires that the various laboratories have similar capabilities in examining particular marker classes. When allozymes were used by most laboratories, standardization of loci could easily be accomplished by the use of common electrophoretic conditions, staining methods and locus-protein interpretations. However, with the advent of DNA technologies, laboratory protocols increasingly depend on the acquisition of costly analytical instruments, such as automated sequencers, to produce genotypic data. Most governmental laboratories charged with the use of genetics in management are able to acquire or have access to equipment to standardize markers among laboratories.

When cooperating laboratories use the same class of markers (*e.g.* allozymes, microsatellites or single nucleotide polymorphisms), standardization of a common set of loci among labs can be achieved through collaboration. As new technologies appear, however, some labs may adopt methods not implemented in other labs. The development of a coast-wide database, in these circumstances, depends on the adoption of the new methods in other labs or the sharing of tissues to extend the range of geographic data for the new marker.

Common nomenclature among laboratories for corresponding allelic states

The second criterion requires that laboratories standardize the nomenclatures of allelic states. Virtually no standardization is required for nucleotide sequences, as only four easily identified nucleotide states are possible. While sequence data are ideal for many applications, they are costly to produce and greatly limit the numbers of individuals and populations that can be reasonably analyzed. The use of single nucleotide polymorphisms is attractive because at least one nucleotide state defines a standard genotype and avoids the need for allelic standardization among laboratories.

As previously with allozymes, the standardization of microsatellite datasets among laboratories requires comparisons of genotypic voucher samples on each analytical platform or the use of allelic ladders (LaHood *et al*. 2002). Different models of automated sequencers, or even the same model in the same laboratory, can produce different electrophoretic mobilities for the same size allele (see Moran *et al*. 2006). The electrophoretic properties of slab gels often differ from capillary tubes so that the same sized microsatellite fragment may have different mobilities in different instruments. Additionally, some alleles deviate from the expected repeat sizes of variable motif, showing apparent sizes that are inconsistent with the repeat motif. The pooling of these alleles of similar sizes must be agreed upon for datasets to be compatible.

Issues 1 and 2 can be resolved by active collaboration among laboratories and periodic workshops to standardize the selection of loci and the nomenclature of alleles. Workshops in 1999, 2000 and 2001 were convened and attended by major agencies to discuss these two issues. In past efforts to standardize allozyme markers, progress toward standardized datasets was slow and incremental over several years, except when agency management directives provided 'specific and immediate motivation' (Moran *et al*. 2006; Seeb *et al*., in press).

Agreed upon sampling of important contributing spawning populations

A third issue involves the standardization of geographical sampling effort. A coordinated dataset of baseline populations requires the same geographical resolution of spawning populations in different regions. As databases expand geographically or are merged with other regional databases greater diversity is encountered in allelic size and may present problems for multiplex analysis of different microsatellite loci on the same electrophoretic system. Greater geographical sampling may compromise the utility of some microsatellite loci as new alleles may produce complex allele frequency distributions that complicate the identification of alleles. Greater allelic diversity is also likely to include null microsatellite alleles, in part due to the failure of polymerase chain reactions (PCR) to amplify a target fragment. Standardization of microsatellite alleles may be more difficult on a broad geographical scale for some species because of these complicating factors, and recommendations must include costs and benefits of sampling at various spatial scales.

Applications of SNPs over large distances may be confronted with other problems. Marker development and sampling strategies are usually shaped by problems under the jurisdictions of regional management agencies. While allelic identification among laboratories may not be

problematic for SNPs, SNP polymorphisms identified in one region may not be present in another region. For example, SNP polymorphisms developed for Alaskan populations may be useful for differentiating Asian populations from North American populations, but may be less informative within Asia.

Use of a common set of statistical procedures

A fourth issue concerns the consistency of statistical analyses among laboratories. One consideration is the identification of genetically discrete populations. Detecting population differences depends on the geography of sampling and on statistical power for finding allele-frequency differences, which is influenced by both sample size and the particular approach to probability adjustment. Sampling design may also influence inferences about population structure as salmon populations can be resolved temporally by run or spawning time and by geography, often on small spatial scales. In addition to the completeness of a population data baseline, the results of mixed-stock analyses depend on the timing and sizes of samples from ocean or river mouth harvests, on reporting aggregations of baseline populations, and on the statistical method used to estimate the composition of the mixture.

Access to data

The foregoing considerations set the stage for the sharing of genetic data to conduct mixed-stock analyses of fishery harvests and to infer ocean abundances and migratory pathways of particular populations. Genetic data now play a fundamental role in the management of salmon populations by federal, state, and tribal agencies. The distribution of current, but often unpublished, data is vital to these efforts. Requests for information may include tissue samples for additional analyses, genotypic or allele frequency data, summary statistics or draft reports. Although funding from federal agencies often comes with agreements on data sharing, genetic databases are usually constructed over several years with multiple sources of funding. Data sharing directives in the USA are embodied in the Freedom of Information Act of 1986 (FOIA), guidelines from the Department of Justice and the Office of Management and Budget, court judgments and executive orders (Moran *et al*. 2006). FOIA requests for the release of genetic data in a timely manner, however, can be impeded by three exemptions: 1) confidential trade information, 2) pre-decisional legal deliberations, and 3) criminal investigations. To date, no court deliberations have commented specifically on the use of FOIA to obtain genetic data (Moran *et al*. 2006). Data-sharing agreements between governmental agencies within a particular country, however, have limited value for facilitating data sharing between agencies in different countries.

META-DATABASE

The first step toward facilitating the easy distribution of data is to establish a web-based electronic ‘meta-database’ that would be easily accessible to stakeholders and management. The primary function of this database would be to catalogue existing primary genetic data (markers, sample dates and sampling localities), biological information (population profiles) and biological materials (tissues, otoliths and scales) that can be used for genetic analysis. The mandates of the present workshop provide the impetus for the construction of such a database. A meta-database,

however, would be logistically complex and would require continuing support to maintain as new information became available. A similar genetic meta-database is being established by ICES for commercially important species in the North Atlantic (ICES 2007).

Several benefits would accrue to stakeholders and users. A meta-database would allow researchers and fishery managers to immediately gather relevant information on databases and researchers for a particular fishery management problem. A meta-database would also improve the designs of research projects and sampling. This database might include the following:

- Existing allozyme, mtDNA, microsatellite, SNP, and EST datasets and where they are located;
- Existing collections of historical biological material that could be used to extract DNA. Archived scales and otoliths can be used to estimate allele frequencies in past populations;
- List of past and current genetics projects, including laboratory location, researcher names and the natures of the projects;
- Profiles and contact information of active researchers working on the genetics of salmon.

This database would provide ready access to information on experts and on geographic areas where data are available. The development of a frequently updated meta-database would promote communication between geneticists and between geneticists and other researcher and managers. Such a database would also help to reduce the overlapping of sampling effort and encourage collaborations and lead to more efficient research efforts. The present state-of-the-art software could be used to make the meta-database portable, so that the responsibility of maintaining the database could be rotated periodically among agencies. The development of a meta-database of information on genetic markers, regional datasets and researchers, would be the first step in establishing a central database containing raw or summary data that could be used by fishery managers.

DATA SHARING

Several stumbling blocks appear to prevent the easy exchange of data among research agencies. A fundamental concern of many researchers is the protection of proprietary information for use in scientific publications. An agreement for the use of current, unpublished genetic data would have to be made between researchers producing genetic data and agency personnel wanting to use the information for management or conservation evaluations. The use of unpublished data by agencies has longstanding precedents in the writing of status reviews on threatened and endangered species by the US Fish and Wildlife Service and NOAA Fisheries (Waples 1991, and numerous status reviews).

Another facet of this problem is that university researchers are disinclined to maintain large databases or to routinely analyze large numbers of samples for management. The development of comprehensive databases usually falls under the mandates of state, national, and international

fishery management agencies. Presently, most genetic data for Pacific salmon in the North East Pacific are held by governmental agencies to assist in their management obligations. Nevertheless, other databases have been generated or are maintained by tribal or university researchers. Ultimately, the easy sharing of data depends on the goodwill and cooperation of personnel at these laboratories.

Agencies may hesitate to share data for fear that some interpretations of a dataset may not prove beneficial to a particular stakeholder's take of the harvest. Differences in interpretation can potentially arise from the use of different statistics or the inclusion of some samples but not others in a database used for mixed stock analysis in areas including fish originating from different jurisdictions. Such differences must be negotiated in the light of the best possible use of data and statistics.

Motivation for sharing and building a standardized dataset for each species of interest to the Pacific Salmon Commission arises from two sources. As scientists, laboratory directors are interested in researching and testing hypotheses that illuminate the sources of genetic population structure in view of historical and contemporary evolutionary and ecological processes. A large body of literature based on genetic data for Pacific salmon has been the cornerstone in fish biology and fisheries management circles for understanding the effects of harvests and climate change on fish populations. No other group of fishes has been examined with genetic markers to the same extent as Pacific salmon.

Laboratory directors in management agencies are also motivated by the mandates of their agencies to manage natural resources as sustainably as possible. The mandate of the Pacific Salmon Commission is to provide management information in areas where fish from different national jurisdictions potentially mix. While these management problems may be limited to trans-boundary areas, the management of these areas often depends on coast-wide databases of populations potentially contributing to harvests in trans-boundary areas. Hence, integrated genetic datasets are all the more important. While funding agencies may impose data-sharing requirements on researchers, laboratories generally received support from several in-house and agency sources, each of which may have different data-sharing mandates. As agency laboratories are part of a hierarchy, the ultimate responsibility for data sharing lies with the administrations of these agencies. When problems arise among laboratories, cooperation may have to be implemented by memoranda of agreements that clearly outline lines of responsibility and how shared data can be used.

CITATIONS

ICES. 2007. Report of the working group on the application of genetics in fisheries and mariculture (WGAGFM), 19–23 March 2007, Ispra, Italy. International Council for the Exploration of the Sea.

LaHood, E.S., Moran, P., Olsen, J., Grant, W.S., Park, L.K. 2002. Microsatellite allele ladders in two species of Pacific salmon: preparation and field-test results. *Molecular Ecology Notes* 2: 187–190.

Moran, P., Teel, D.J., LaHood, E.S., Drake, J., Kalinowski, S. 2006. Standardising multi-laboratory microsatellite data in Pacific salmon: an historical view of the future. *Ecology of Freshwater Fish* 15: 597–605.

Seeb, L.W., Crane, P.A., Kondzela, C.M., Wilmot, R., Urawa, S.N. 2004. Migration of Pacific Rim chum salmon on the high seas: insights from genetic data. *Environmental Biology of Fishes* 69: 21–36.

Seeb, L.W., Antonovich, A., Banks, M.A., Beacham, T.D., Bellinger, M.R., Campbell, M., Garza, J.C., Guthrie III, C.M., Moran, P., Narum, S.R., Stephenson, J.J., Supernault, K.J., Teel, D.J., Templin, W.D., Wenburg, J.K., Young, S.F., Smith, C.T. Submitted. Development of a Standardized DNA Database for Chinook Salmon. *Fisheries*.

Teel, D.J., Van Doornik, D.M., Kuligowski, D.R., Grant, W.S. 2003. Genetic analysis of juvenile coho salmon (*Oncorhynchus kisutch*) off Oregon and Washington reveals few Columbia River wild fish. *Fishery Bulletin* 101: 640–652.

Van Doornik, D.M., Teel, D.J., Kuligowski, D.R., Morgan, C.A., Casillas, E. 2007. Genetic analysis provide insight into early ocean stock distribution and survival of survival of juvenile coho salmon off the coasts of Washington and Oregon. *North American Journal of Fisheries Management* 27: 220–237.

Waples, R.S. 1991. Pacific salmon, *Oncorhynchus* spp., and the definition of "species" under the Endangered Species Act. *Marine Fisheries Review* 53: 11–22.

APPENDIX D. INDIVIDUAL ASSIGNMENTS AND STOCK COMPOSITION ESTIMATES FOR A MIXTURE WHEN SOURCE MARKS ARE NOT DEFINITIVE

Jerome Pella

Post Office Box 210332, Auke Bay, Alaska 99821

INTRODUCTION

Artificial marks, such as coded wire tags and otolith thermal marks, are created to provide definitive source identification for individual fish found in stock mixtures. Unless an infrequent human error is made during application or recovery, the origin of a fish carrying an artificial mark is known with certainty. The usual shortcomings of artificial marks include their expense in application and in determination of the source at recovery. Further, artificial marking is incomplete in scope because neither all the stocks nor all the individuals in the stocks composing mixtures are marked. Natural marks of individuals, such as scale features, parasites, and genotypes, provide less certain source identification. Usually the contributing stocks to a mixture share all the natural marks (*i.e.*, individuals with the various marks are found in all the stocks) but the relative frequencies of the marks differ among stocks (*i.e.*, the proportions of individuals with any one of the various marks differ among the stocks). The advantages of natural marks are complete coverage of all stocks as well as all individuals in the stocks. However, the cost of baseline development (*i.e.*, the initial samples to characterize the distributions of the natural mark among individuals of each stock) may be large and the cost in sampling mixtures and determining the natural marks of mixture individuals may be significant. Although source composition estimation of a mixture and of the origins of individuals in a sample requires more complex methods for natural marks than for artificial marks, appropriate statistical theory and estimation algorithms are well-developed, and software for their implementation is freely available.

ARTIFICIAL MARKS

Artificial marks identify each individual to its source, and so the problem of estimating the sources of individuals in a mixture sample does not apply. For example, if 100% of individuals of each stock are marked, the problem of estimating the source composition of the mixture from its random sample is solved through straightforward application of multinomial sampling theory. Assume that *c* stocks occur in a mixture. The unknown source composition of the mixture $\mathbf{p}$ occurs on the simplex, $S(\mathbf{p}) = \left\{ \mathbf{p} = (p_1, \mathsf{K}\ , p_c)' \right\}$ where $0 \le p_i \le 1,\ i = 1, \mathsf{K}\ , c,$ and $\sum p_i = 1$. The array of stock counts found in a random sample of size *M* from the mixture is denoted as $\mathbf{m} = (m_1, \mathsf{K}\ , m_c)'$, where m_i is the count of individuals from the *i-th* stock and $M = \sum_{i=1}^{c} m_i$. The multinomial probability function that describes the sampling variation is

$$\text{Prob}(\mathbf{m}) = \frac{M!}{m_1! \mathsf{K}\ m_c!} p_1^{m_1} \mathsf{K}\ p_c^{m_c}.$$

The obvious estimator of the stock composition of the mixture is the observed stock composition of the sample itself. If frequentist methods of estimation are used, the maximum likelihood estimate (MLE) of the stock composition of the mixture is simply this observed stock composition $\hat{\mathbf{p}} = \left(m_1 / M, \mathsf{K}\ , m_c / M \right)'$, it is unbiased in that the average value equals the actual

composition over repeated sampling, $E(p_i) = p_i,\ i = 1,\mathrm{K}\ ,c$, and its estimated covariance matrix is $\hat{\Sigma}_{\hat{\mathbf{p}}} = \left[\hat{\sigma}_{ij}\right]$, where $\hat{\sigma}_{ii} = \hat{p}_i(1-\hat{p}_i)/M$ and $\hat{\sigma}_{ij} = -\hat{p}_i\hat{p}_j/M$ for $i, j = 1,\mathrm{K}\ ,c, i \neq j$. The MLE estimator of the stock composition is sensible provided no additional information besides the mixture sample is available by which to estimate the stock composition.

NATURAL MARKS

The source identity of an individual is almost never certain from its natural marks and so both the sources of individuals and the stock composition of the mixture must be estimated. The duality of the estimation problem—individual sources and stock composition—is used hereafter to motivate the various solutions that have been developed. Here we introduce the methods in order of their increasing strength and suitability.

Classical Individual Assignments Method

The classical individual assignments method is an ostensibly reasonable approach to the dual estimation problem and comprises two steps that are applied just once to the mixture sample. First, assign the individuals to their sources based on their marks and the relative frequencies of their marks in a set of baseline samples from all the possible source stocks. Second, estimate the mixture composition from the assignments using the multinomial theory described under the preceding section. At the second step, the assignments are treated as accurate, and the probable errors in the assignments are ignored. Quite likely, early workers in scale pattern analysis during the 1950s used this method before the statistical adjustments for assignment errors by Worlund and Fredin (1962), Cook and Lord (1978), Pella and Robertson (1979), Millar (1987) and Wood *et al*. (1987) (see summary by Pella and Masuda 2005). The assignment errors were well known to cause both bias and overstated precision in estimated mixture composition. Nonetheless, many geneticists have also used the method in more recent times (Banks and Eichert 2000, Potvin and Bernatchez 2001) that were evidently unaware of the statistical adjustments by scale pattern analysts.

In addition to neglecting the effects of assignment errors on the mixture composition estimates, the assignment rule commonly used in the method is inferior to another well-known rule that has lower expected error rate for nearly every possible mixture provided, *and here is the crux*, that the mixture composition is specified. Practitioners usually assign an individual with measurement vector **X** (this could be any of the following variable subsets for an individual fish as well as their combination: scale characters, morphometric measurements, binary parasite occurrence indicators, and multilocus genotype indicators) to the source population for which the measurement is most common or frequent (the rule will be termed the maximum frequency or MAF rule). Specifically, the individual is assigned to the stock *i** for which $\hat{f}_{i*}(\mathbf{X}) = \max_i \left\{\hat{f}_1(\mathbf{X}),\mathrm{K}\ ,\hat{f}_c(\mathbf{X})\right\}$, where $\hat{f}_i(\mathbf{X})$ is the estimated relative frequency (probability function if **X** is discrete, or probability density if **X** is continuous) of individuals with measurement **X** in the *i-th* stock. The underlying relative frequency, $f_i(\mathbf{X})$, is unknown and so the estimated value from the baseline samples is used in its place. Notice that the relative

frequency of individuals with measurement **X** in the mixture is $f(\mathbf{X}) = \sum_{i=1}^{c} p_i f_i(\mathbf{X})$, where p_i is the proportion of the mixture from the *i-th* stock, and that the fraction of such individuals contributed by the *i-th* stock is the ratio specific to **X (**called the posterior source probability of that stock), $P(i \mid \mathbf{X}) = p_i f_i(\mathbf{X}) \Big/ \sum_{j=1}^{c} p_j f_j(\mathbf{X})$, $i = 1, \ldots, c$. What better guess for the source of an individual with measurement **X** than the stock contributing the most individuals with the measurement? In fact, assignment errors are minimized by using this so-called Bayes' classifier in which the individual with measurement **X** is assigned to the stock for which the posterior source probability is highest (called the maximum *a posteriori*, or MAP rule), that is, to the *i*-th* stock if $P(i^* \mid \mathbf{X}) = \max_i \{P(1 \mid \mathbf{X}), \ldots, P(c \mid \mathbf{X})\}$. Notice that if the stocks contribute equally to the mixture so that $p_1 = p_2 = \ldots = p_c$, only then do the MAF and MAP rules agree. In fact, the stock proportions cancel from the formula for the posterior source probabilities so that the analyst appears to be relieved of having to provide these values. A further justification for using the MAF rule may be an apparent lack of information about the mixture composition. In fact, the very reason for performing the assignments is usually to estimate the unknown composition. Although this argument for the MAF rule claiming ignorance seems reasonable at first glance, it fails to convince after the assignments are completed and some knowledge regarding composition becomes available. In general, the estimated composition from the assignments will differ from the assumed equal composition. If the classical individual assignments method were trusted to produce a more accurate composition estimate than the initial equal proportions assumption, why not substitute the better estimate for unknown **p** into the posterior source probabilities, and repeat assignments with the superior MAP rule? In fact, what about repeating this process to convergence in the estimate of **p** if possible? As we emphasized earlier, the superior MAP rule is available provided the mixture composition can be specified. An iterative series of assignments using the MAP rule with the resulting mixture composition estimates to restart the rule provide the recipe. The only inconvenience is that the assignments and estimation need to be done again and again, but that is what computers do so well. The approach is well grounded in sound statistical theory. Also, suppose that none of the posterior source probabilities for an individual is large relative to the others. Should the individual be assigned to a single stock, or would it be better to assign it fractionally to the possible stocks in proportion to the posterior source probabilities? Why not fractionally assign every individual regardless of the relative magnitudes of the posterior source probabilities? Two modern valid approaches are discussed next in the context of genetic marks in which these ideas are completed. The general approach is termed mixture modeling, under which two methods are important, conditional maximum likelihood and Bayesian. To see the parallel developments for non-genetic characters using discriminant analysis, see the review by Pella and Masuda (2005).

Conditional Maximum Likelihood Method of Mixture Modeling

The first valid estimation method developed for the stock composition of a genetic mixture is called the conditional maximum likelihood method (Fournier et al. 1984, Millar 1987, Pella and Milner 1987). The term "conditional" refers to the fact that the genetic parameters are estimated using so-called baseline samples from the contributing stocks and the likelihood function is

maximized only with respect to the unknown stock proportions. The genetic parameters for the *i-th* stock are denoted by $\mathbf{Q}_i = (q_{ihj})$, $h = 1,\mathsf{K}\ L;\ j = 1,\mathsf{K}\ ,J_h$ and refer to the allele relative frequencies at the *L* loci defining the multilocus genotypes of individuals ($0 \le q_{hj} \le 1,\ \sum_{j=1}^{J_h} q_{hj} = 1$). The relative frequencies of multilocus genotypes in any stock are computed from estimates of allele relative frequencies under Hardy-Weinberg and linkage equilibrium conditions. Let the multilocus genotype of the *m-th* mixture individual be denoted by $\mathbf{X}_m = (\mathbf{X}_{m1},\mathsf{K}\ ,\mathbf{X}_{mL})$, where $\mathbf{X}_{mh} = (x_{mh1},\mathsf{K}\ ,x_{mhJ_h})$ is the vector of allele counts for the individual at the *h-th* locus. Note that each individual has 2 alleles per locus giving $x_{mh+} = \sum_{j=1}^{J_h} x_{mhj} = 2$. Then the relative frequency of the genotype of the *m-th* individual in the *i-th* stock is $f(\mathbf{X}_m;\mathbf{Q}_i) = \prod_{h=1}^{L} 2^{1-\delta_h(\mathbf{X}_m)} \prod_{j=1}^{J_h} q_{ihj}^{x_{mhj}}$, where $\delta_h(\mathbf{X}_m) = 1$ if the individual is homozygous, and equals 0 if it is heterozygous, for locus h. These Hardy-Weinberg and linkage equilibrium conditions are quite plausible for large panmictic populations that do not exchange immigrants. The estimation of the allele relative frequencies $\mathbf{Q} = (\mathbf{Q}_1,\mathsf{K}\ ,\mathbf{Q}_c)$ is considered shortly and for now we use the unknown value recognizing that an estimate is substituted for computations.

The probability of the genotypes, $\mathbf{X}_1,\ldots,\mathbf{X}_M$, observed in a random sample of *M* individuals from the mixture is

$$\text{Prob}\ (\mathbf{X}_1,\mathsf{K}\ ,\mathbf{X}_M) \propto \prod_{m=1}^{M}\left(\sum_{i=1}^{c} p_i f(\mathbf{X}_m;\mathbf{Q}_i)\right).$$

It can be shown that the maximizing value of **p** given the genotypes can be obtained by iteratively ($t = 1,2,\ldots,T$) solving the equation system,

$$p_1^{(t)} = \frac{1}{M}\sum_{m=1}^{M} \frac{p_1^{(t-1)} f(\mathbf{X}_m;\theta_1)}{\sum_{i=1}^{c} p_i^{(t-1)} f(\mathbf{X}_m;\theta_i)} = \frac{1}{M}\sum_{m=1}^{M} P^{(t-1)}(1 \mid \mathbf{X}_m)$$

M

$$p_c^{(t)} = \frac{1}{M}\sum_{m=1}^{M} \frac{p_c^{(t-1)} f(\mathbf{X}_m;\theta_c)}{\sum_{i=1}^{c} p_i^{(t-1)} f(\mathbf{X}_m;\theta_i)} = \frac{1}{M}\sum_{m=1}^{M} P^{(t-1)}(c \mid \mathbf{X}_m).$$

An arbitrary value for $\mathbf{p}^{(0)}$ with positive components of unit sum can be substituted into the right hand side of the equation system to start the iteration. The composition for $\mathbf{p}^{(1)}$ results on the left hand side, and it is substituted into the right hand side again. The process is continued to convergence, which can be judged to have occurred by various criteria. Typically convergence is assumed when the changes between successive estimates of **p** become arbitrarily small. This method of computing the conditional maximum likelihood estimate $\hat{\mathbf{p}}$ is called the EM algorithm. Although the conditional MLE for **p** by the EM algorithm may not be as fast to

compute as by other algorithms (Pella *et al*. 1996), it has the advantages of simplicity and the guarantee to converge with monotonic increase in likelihood function values during the search. Another advantage is the simple and intuitive interpretation of what is being done to compute the conditional MLE. Notice from the equation system that at convergence each individual has a unit value that is divided up among the possible stocks in proportion to its posterior source probabilities, and that the stock composition estimate is the arithmetic average of these posterior source probabilities. Therefore, instead of assigning the entire individual to a stock as though one knew its source, the individual is assigned fractionally in proportion to measures of our belief of its sources.

The estimation of the allele relative frequencies **Q** among loci and stocks has become more complex as choice in genetic marks changed. Initially, allozymes were available and these loci typically had only a few alleles per locus. Therefore, maximum likelihood estimation of the allele frequencies was based on the standard multinomial sampling model described under the section "Artificial Marks". Random samples of fish from escapements, called baseline samples, were assayed for their genotypes, *i.e.*, the pairs of alleles per fish. The maximum likelihood estimates of the $\mathbf{Q}_i$ are the observed allele compositions of the combined alleles of the baseline sample from each stock and locus. More recently, microsatellite loci were used for which the number of different alleles among stocks could be large (as many as 50 or even more) and many of these alleles were rare or in low relative frequency in stocks. The result was that sampling zeros were presumably common in the baseline samples and this became very problematic to stock composition estimation. If maximum likelihood was used to estimate the allele relative frequencies from the baseline samples, stocks whose baseline samples had sampling zeros were necessarily eliminated as potential sources for mixture individuals having the corresponding alleles in their genotypes. Other stocks without sampling zeros became the candidates, and with far greater posterior source probabilities than should have occurred. Because certainty about the absence of a rare allele is not possible with limited sample size from large populations, stock composition estimation from microsatellites performs better if the possibility is maintained that any baseline stock could be the source of any mixture individual. This possibility is achieved through use of Bayesian methods that in the present context provides a probability distribution for allele relative frequencies of **Q** rather than a point estimate and associated measure of variation. Under this Bayesian probability distribution, every allele is potentially present in every stock. Were a point estimate needed, the location parameter, *e.g.*, mean or median, of the Bayes distributions of components of **Q** should suffice.

The recommended method of estimating the sampling variation in the conditional MLE, $\hat{\mathbf{p}}$, is by bootstrap resampling. The reason that asymptotic methods are not useful for this purpose is they are inaccurate and overestimate the uncertainty. Asymptotic methods are inaccurate because they depend on an assumption, nearly always violated, that the distributions of composition estimates do not encounter the boundaries for the unknown proportions ($0 \le p_i \le 1,\ i = 1,\mathsf{K}\ ,c,\ \sum_{i=1}^{c} p_i = 1$).

In the bootstrap method, the mixture and baseline samples are sampled with replacement to generate random analogs of the same size. The allele relative frequencies **Q** are estimated by either maximum likelihood from the bootstrap baseline samples (appropriate if none of the

alleles is rare), or by a single draw from the Bayes distribution of **Q** (appropriate when some alleles are rare). Then the conditional maximum likelihood estimate of **p** is computed for the bootstrap mixture sample using the EM algorithm, for example. A large number of repetitions of this process (*e.g*., 1000 times) generates an empirical bootstrap distribution from which the mean and lower $\alpha/2 \cdot 100$-*th* percentile and upper $(1-\alpha/2) \cdot 100$-*th* percentile provide a point estimate and symmetric $(1-\alpha) \cdot 100\%$ confidence bounds.

Software to perform the conditional maximum likelihood estimation method and bootstrap resampling of baseline and mixture samples is freely available from two sources. The earliest software is the program called Statistical Package for Analyzing Mixtures (SPAM) (Debevec *et al*. 2000). SPAM originally did not include the option during bootstrap resampling of using the Bayesian posterior for baseline allele relative frequencies. Meanwhile, Kalinowski (2003) developed the program called Genetic Mixture Analysis (GMA) that did. Later, SPAM was updated to include the option (Alaska Department of Fish and Game 2003).

Bayesian Method of Mixture Modeling

The Bayesian approach has been extended to include estimation of both the stock composition **p** and the allele relative frequencies **Q** (Pella and Masuda 2001). The combination of a Bayesian analysis for **Q** and a maximum likelihood approach for **p** described above is a peculiarity in that frequentist and Bayesian statisticians view any estimation problem as mirror images. Under Bayesian methods, data are considered fixed and unknown parameters are considered random variables. Under frequentist methods such as maximum likelihood, data are considered random and parameters are fixed. Although the two schools have long debated the validity of their approaches, the more recent Bayesian methods have gained favor by more exact modeling of complex problems made possible by availability of greater computing power. The premise of the Bayesian method for estimation of a collection of unknowns, say $\Theta = (\mathbf{p}, \mathbf{Q})$, is that information exists about Θ before a sample is drawn and data **Y** become available. The information is provided in the form of a prior probability distribution, in the present problem by $\pi(\mathbf{p}, \mathbf{Q}) = \pi(\mathbf{p})\pi(\mathbf{Q})$. That is, the prior information about the stock proportions, **p**, and genetic parameters, **Q**, is assumed to be statistically independent so that their joint prior probability distribution equals the product of their separate prior distributions. Uninformative priors are chosen for both **p** and **Q** so that the data "do the talking". The information in the data **Y** is combined with that of the prior by integration of the product of the prior, $\pi(\mathbf{p}, \mathbf{Q})$, and the likelihood of the data, $\pi(\mathbf{Y} \mid \mathbf{p}, \mathbf{Q}) \propto \prod_{m=1}^{M} \left[\sum_{i=1}^{c} p_i f(\mathbf{X}_m \mid \mathbf{Q}_i) \right]$, to produce the posterior distribution, $\pi(\mathbf{p}, \mathbf{Q} \mid \mathbf{Y}) \propto \int \pi(\mathbf{p}, \mathbf{Q})) \cdot \pi(\mathbf{Y} \mid \mathbf{p}, \mathbf{Q}) d\mathbf{p} d\mathbf{Q}$. The posterior distribution summarizes the knowledge and uncertainty about the unknowns. Pella and Masuda (2001) chose so-called conjugate priors for both **p** and **Q** so that the Bayes posterior distribution for the unknowns has an explicit solution. The prior for **p** is the Dirichlet density function, which is defined on the stock composition simplex, $S(\mathbf{p}) = \left\{ \mathbf{p} : 0 < p_i < 1, \sum_{i=1}^{c} p_i = 1 \right\}$. The density function is

parameterized by $\alpha = (\alpha_1, \mathrm{K}, \alpha_c)$ and is $\pi(\mathbf{p} \mid \alpha) = D(\mathbf{p} \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{c} \alpha_i\right)}{\prod_{i=1}^{c} \Gamma(\alpha_i)} \prod_{i=1}^{c} p_i^{\alpha_i - 1}$. Pella and Masuda (2001) set the $\alpha_i = 1/c,\ i = 1, \mathrm{K}, c$, which has the desired effect of providing low information about **p**: this prior information is equivalent to adding a single individual to the mixture sample and specifies that the prior contributions from the source stocks to the mixture are equal. If the source identities of the *M* individuals in the random multinomial mixture sample were available and the array of stock counts is denoted by $\mathbf{Z} = (z_1, \mathrm{K}, z_c)$, where $\sum_{i=1}^{c} z_i = M$, the posterior distribution is the Dirichlet density function, $\pi(\mathbf{p} \mid \mathbf{Z}) = D\left(\mathbf{p} \mid \mathbf{Z} + \frac{1}{c} \cdot \mathbf{1}\right) = D\left(\mathbf{p} \mid \mathbf{Z} + \frac{1}{c} \cdot (1, \mathrm{K}, 1)\right)$. The means, variances, and covariances of this distribution are given by

$$E(p_i \mid \mathbf{Z}) = (z_i + c^{-1}) / (M + 1),\ i = 1, \mathrm{K}, c$$

$$\mathrm{var}(p_i \mid \mathbf{Z}) = \left[(z_i + c^{-1})(M + 1 - (z_i + c^{-1})) \right] / \left[(M + 1)^2 (M + 2) \right],\ i = 1, \mathrm{K}, c$$

$$\mathrm{cov}(p_i, p_j \mid \mathbf{Z}) = -\left[(z_i + c^{-1})(z_j + c^{-1}) \right] / \left[(M + 1)^2 (M + 2) \right],\ i, j = 1, \mathrm{K}\ c$$

As M becomes large, these values for the posterior distribution of **p** agree closely with the MLE estimates of the corresponding values from the multinomial probability function (substitute components of **Z** in place of **m** into the multinomial formulas under "Artificial Marks"), so that the Bayesian posterior distribution will be a reasonable description of knowledge and uncertainty for both Bayesian and frequentist statisticians. Of course, the actual counts of individuals by source stock are unknown and what the Bayesian method does to accommodate this uncertainty is described shortly.

The remaining unknowns are the allele relative frequencies **Q** at *H* different loci among the *c* stocks. Two information sources about **Q** are available: the baseline samples, and the mixture sample. In contrast to the conditional maximum likelihood method, which uses only the baseline samples for estimating **Q**, the Bayesian method extracts the information about **Q** from both. First, a separate Bayesian analysis is performed with the baseline samples to develop the baseline posterior distribution for **Q**. Second, the baseline posterior distribution for **Q** becomes the mixture prior for **Q** to be updated during the mixture sample analysis. The baseline sample for the *h-th* locus from the *i-th* stock is viewed as a random draw from the multinomial probability function with J_h different alleles possible at the locus. In practice, the value for J_h is the number of different alleles at the locus observed among the baseline stocks. Again, the Dirichlet prior probability density has been used to describe knowledge and uncertainty in **Q**. Two specifications for the parameters of the Dirichlet prior have been used. The most straightforward specification (Rannala and Mountain 1997) is the analog to that described above for stock composition **p**, but now applied to each of the $c \cdot H$ unknown allele relative frequency arrays denoted by $\mathbf{q}_{ih} = (q_{ih1}, \mathrm{K}, q_{ihJ_h}),\ i = 1, \mathrm{K}, c,\ h = 1, \mathrm{K}, H$ (see Kalinowski 2003). To distinguish

these prior parameters from those used for **p**, they will be denoted by $\beta_{ih} = \left(\beta_{ih1}, \mathsf{K}\ , \beta_{ihJ_h}\right)$, where $\beta_{ihj} = J_h^{-1}$ if the *h-th* locus has J_h alleles among the baseline stocks. The other specification by Pella and Masuda (2001) chooses the prior parameters for a locus to be proportional to a baseline center of allele relative frequencies. The baseline center is the unweighted arithmetic average of the observed allele relative frequencies at the locus among the stocks in the baseline. The value for the constant of proportionality of each locus is chosen to minimize the sum of squared deviations between the observed allele relative frequencies and their posterior mean. With this definition for the prior parameters, the prior mean equals the baseline center, and the posterior mean for any stock is the weighted average of its observed allele relative frequencies and the baseline center with the weights equal to simple ratios involving the baseline sample sizes and the constant of proportionality. The method is called pseudo-Bayes because it hedges by using the baseline samples both to choose the prior parameters and to evaluate the posterior distribution for the allele relative frequencies.

The Bayesian approach to describing the knowledge and uncertainty in the unknowns for complex problems is to generate a very large number of samples from their posterior distribution. Then summary measures of the posterior distribution, such as for location (mean, median, mode) and variation (standard deviation and various quartiles) can be computed from the samples with ignorable sampling error. In particular, the Bayesian method for stock mixtures used by Pella and Masuda (2001) is called a Markov chain Monte Carlo (MCMC) method because the samples for the unknowns are generated sequentially with each depending on the immediately preceding sample. The Bayesian method for stock mixtures also uses the data augmentation algorithm in which additional random observations, namely the unknown and purported sources of mixture individuals, are generated to greatly simplify estimation. At the *k-th* sample of the unknowns, let the current values from the posterior of **p** and **Q** be denoted by $\mathbf{p} = \mathbf{p}^{(k)}$ and $\mathbf{Q} = \mathbf{Q}^{(k)}$. The data augmentation algorithm has two steps:

1. Draw a random stock source of each mixture individual, $\mathbf{z}_m^{(k)} = \left(z_{m1}^{(k)}, \mathsf{K}\ , z_{mc}^{(k)}\right)$, where $z_{mi}^{(k)} = 1,\ i = 1, \mathsf{K}\ , c$, if the source of the *m-th* individual is the *i-th* stock, and $z_{mi}^{(k)} = 0$ otherwise. The stock source is drawn with *c* probabilities proportional to the posterior source probabilities based on the genotype and current values of **p** and **Q**.
2. Draw new values $\mathbf{p} = \mathbf{p}^{(k+1)}$ and $\mathbf{Q} = \mathbf{Q}^{(k+1)}$ from their respective posterior densities given the mixture sample genotypes, **X**, baseline samples for allele relative frequencies, **Y**, and the stock identities at step 1, $\mathbf{Z}^{(k)}$.

The posterior distribution for **p** is obtained by updating the Dirichlet prior for **p** with the assigned stock identities for the mixture individuals,
$\pi\left(\mathbf{p} \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}^{(k)}\right) = D\left(\mathbf{p} \mid \frac{1}{c} + \sum_{m=1}^{M} z_{m1}^{(k)}, \mathsf{K}\ , \frac{1}{c} + \sum_{m=1}^{M} z_{mc}^{(k)}\right)$. Notice that each updated Dirichlet parameter for a stock equals the sum of its prior parameter and the total number of mixture individuals assigned to the stock at the preceding sample in the chain. The posterior density for $\mathbf{Q}_i$ is
$\pi\left(\mathbf{Q}_i \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}^{(k)}\right) = \prod_{h=1}^{H} D\left(\mathbf{p} \mid \beta_{h1} + y_{ih1} + \sum_{m=1}^{M} z_{mi}^{(k)} x_{mh1}, \mathsf{K}\ , \beta_{hJ_h} + y_{ihJ_h} + \sum_{m=1}^{M} z_{mi}^{(k)} x_{mhJ_h}\right)$. Notice that each

updated Dirichlet parameter for an allele in the *i-th* stock equals the sum of its prior parameter, the count of the allele in the baseline sample, and the count of the allele for mixture individuals assigned to the *i-th* stock at the preceding sample in the chain.

The data augmentation algorithm generates a chain of samples from the posterior distribution for **p** and **Q**. However, the early samples in the sequence are influenced by the values chosen for **p** and **Q** to begin computations. Early burn-in samples need to be discarded, and a sufficiently large number of subsequent samples must be drawn to describe the posterior distribution accurately. Care is needed to assure that convergence has occurred. Recommended practice is to run several independent chains, ideally a chain for each reporting group, with dispersed starting points to reduce the possibility that a chain is accepted as representative of the posterior distribution before having converged. Usual starting points of the chains are mixture compositions **p** for which each reporting group is preponderant and each member stock contributes equally to the group.

To determine the necessary chain lengths, an iterative scheme of processing pilot chains with the Raftery-Lewis convergence diagnostic (Raftery and Lewis 1996) is applied to each sequence of reporting group proportions (see Pella and Masuda 2001). To monitor convergence, the Gelman-Rubin shrink factor (Gelman and Rubin 1992) is computed for the mixture proportion from each reporting group. After convergence of chains is verified, the MCMC samples after burn-in are combined across chains. Various statistics of the pooled chains (equivalent to parameters of the posterior distribution given the large samples) such as means, standard deviations, and various percentiles (2.5, 5.0, and 97.5) are computed for the reporting group proportions. Also reported for each mixture individual are the chain average relative frequencies of assignment to each of the baseline stocks, which are the averages of the posterior distributions for posterior source probabilities and can be used to assign individuals to their sources by the MAP rule. Pella and Masuda (2001) provide the implementing software (see the Fortran program BAYES on the Auke Bay Laboratory website) for all computations. More recently, the algorithms have been reprogrammed at the Pacific Biological Station as the C program cBAYES, which is available at their website.

DISCUSSION AND SUMMARY

Artificial marks are definitive of the source for individuals, and estimation of mixture composition is straightforward under the multinomial sampling model. Natural marks differ in their distributions among stocks, and the sources of individuals cannot be ascertained with certainty. This document describes various solutions to the dual problem of estimating both the stock sources of individuals and the mixture composition from the natural marks in a random sample of the mixture. In this discussion, we restrict the natural marks to be genetic even if the term is broader. Two general approaches to the dual problem have been used: classical individual assignments and mixture modeling. The classical individual assignments method couches the solution in terms evocative of hypothesis testing (Banks and Eichert 2000, Cornuet et al. 1999, Luikart and England 1999, Paetkau *et al*. 1995). The individuals are assigned to the baseline stock in which the relative frequency of their genotype is the greatest (maximum frequency or

MAF rule). These assignments are performed once and no further use from learning about the mixture composition is made. The proponents call this an assignment test. The assignment tests partition the possible genotypes into assignment classes that have a one-to-one correspondence with the baseline stocks. Each assignment class is composed of those genotypes whose relative frequencies are greatest in its associated stock. Regardless of the assignments made of other mixture individuals, each mixture individual is assigned to the stock corresponding to its assignment class based on its genotype. Because any of the genotypes are estimated as present in any of the stocks, sources of individuals are uncertain and misclassifications occur. After the individuals have been assigned to the stocks based on their genotypes, their genotype relative frequencies among stocks are not used further and their uncertain stock sources are treated as if known and correct. The proportions assigned to the possible sources estimate the source composition. This apparent stock composition from the assignments is biased by misclassifications and the precision of the composition estimate is overstated because the uncertain stock sources of the individuals are treated as known (*e.g.*, see pp. 532-536 of Pella and Masuda 2005).

In the mixture modeling approach, the genotypes in the mixture sample are viewed as having been drawn from the baseline stocks with prior probabilities equal to the unknown mixture composition. The relative frequencies of the genotypes in the mixture are written as weighted sums of their relative frequencies in the contributing stocks. The weights are the unknown stock proportions, *i.e.*, the prior probabilities, composing the mixture. The estimation problem is seen to be a decision problem and Bayes' theorem of mathematical statistics is used to provide the posterior source probabilities of each genotype **X**. Two methods of estimating the stock proportions are described: conditional maximum likelihood and a Bayesian method. One algorithm for the conditional maximum likelihood method is used to demonstrate that individuals are assigned fractionally to the source stocks in proportion to their posterior source probabilities. If an individual must be assigned as an entity, the choice should be the source with the largest posterior source probability. The Bayesian method uses a Markov chain Monte Carlo method together with the data augmentation algorithm. The Bayesian method is somewhat more efficient than the conditional maximum likelihood method in that it extracts the information for allele relative frequencies in the source stocks **Q** from not only the baseline samples but also the mixture sample. The steps of the Bayesian algorithm show that the mixture individuals are repeatedly assigned as entities at random to the source stocks with probabilities proportional to the current chain values for the posterior source probabilities. The long run chain average relative frequencies of assignment to the baseline stocks can be used with the MAP rule to choose the source of an individual.

Proponents of the classical individual assignments method often demonstrated the capacity of their method with genetic data by application to either simulated or real mixture samples composed of equal or nearly equal contributions of individuals from known sources (Cornuet *et al*. 1999, Manuel *et al*. 2002, Potvin and Bernatchez 2001). The authors were evidently unaware that they are rigging the circumstances such that the maximum frequency (MAF) rule of the classical individual assignments method is even better than the optimal classifier. Ordinarily, application of the true optimal rule, namely the maximum *a posteriori* (MAP) rule, requires estimation of the unknown stock composition of the mixture, but when the MAF rule is applied, the result is the same as if the MAP rule were provided with the supposedly unknown mixture

composition. Performance of the classical individual assignments method would more convincingly be demonstrated with very uneven mixtures (*e.g*., 100% mixtures in which only one of the baseline stocks is present), but when the genotypes are less powerful for stock identification, the little evidence available shows as expected that the method is inferior to the MAP rule under these conditions.

Koljonen *et al*. (2005) compared both self-assignment tests and independent tests of mixture samples for 26 stocks of Baltic Atlantic salmon. In the self-assignment tests, each of the 26 baseline samples played two roles: first, as one of the 26 baseline samples for estimating the allele relative frequencies in the source stocks, **Q**, and, second, as a 100% pure mixture sample composed entirely of a single stock. Among the 26 self-assignment tests, the average estimated correct stock proportion was 75.1% (range 44%-94%) by classical individual assignments (using program GeneClass of Cornuet *et al*. 1999) versus 97.0%, the average of the 26 Bayesian posterior averages (posterior averages ranged from 91.0%-98.7%) by mixture modeling using the Bayesian method (program BAYES of Pella and Masuda 2001). Self-assignment tests are biased toward better performance than can be expected with independent mixture samples. Two independent pure stock samples were available: 50 hatchery fish from the Neva River in Russia and 56 wild fish from the Tornionjoki River in Finland and Sweden. Program GeneClass identified the correct source of 44 (88%) of the Neva River stock. However, program BAYES estimated the mixture composition to be 97.1% (Bayesian posterior average) from the Neva River stock and when the individuals were assigned by the MAP rule applied to the chain averages of posterior source probabilities for individuals, all (100%) were correctly identified. When the Tornionjoki River wild fish were analyzed, GeneClass correctly identified only 15 of the 56 individuals (26.8%) whereas the BAYES posterior average proportion was 91.0% and BAYES correctly identified 55 of 56 (98.2%) of the individuals correctly using the MAP rule.

CITATIONS

Alaska Department of Fish and Game. 2003. *SPAM Version 3.7: Statistics Program for Analyzing Mixtures*. Anchorage, AK: Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab.

Banks, M., Eichert, W. 2000. WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *Journal of Heredity* 91: 87–89.

Cook, R., Lord, G. 1978. Identification of stocks of Bristol Bay sockeye salmon, *Oncorhynchus nerka*, by evaluating scale patterns with a polynomial discriminant method. *Fishery Bulletin* 76: 415–423.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A., Solignac, M. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* 153: 1989–2000.

Debevec, E., Gates, R., Masuda, M., Pella, J., Reynolds, J., Seeb, L. 2000. SPAM (version 3.2): statistics program for analyzing mixtures. *Journal of Heredity* 91: 509–510.

Fournier, D., Beacham, T., Riddell, B., Busack, C. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. *Canadian Journal of Fisheries and Aquatic Sciences* 41: 400–408.

Gelman, A., Rubin, D.B. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7: 457–511.

Kalinowski, S.T. 2003. *Genetic Mixture Analysis 1.0*. Bozeman, MT: Department of Ecology, Montana State University. Available at http://www.montana.edu/kalinowski

Koljonen, M.-L., Pella, J., Masuda, M. 2005. Classical individual assignments versus mixture modeling to estimation stock proportions in Atlantic salmon (*Salmo salar*) catches from DNA microsatellite data. *Canadian Journal of Fisheries and Aquatic Sciences* 62: 2143–2158.

Luikart, G., England, 1999, Statistical analysis of microsatellite DNA data. *Trends in Ecology and Evolution* 14: 253–256.

Manuel, A., Berthier, P., Luikart, G. 2002. Detecting wildlife poaching: Identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conservation Biology* 16: 650–659.

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

Paetkau, D., Calvert, W., Stirling, W., Strobeck, W. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* 4: 347–354.

Pella, J., Robertson, T. 1979. Assessment of composition of stock mixtures. *Fishery Bulletin* 77: 387–398.

Pella, J., Milner, G. 1987. Use of genetic marks in stock composition analysis. In *Population genetics and fisheries management* (N Ryman, F. Utter, eds), pp. 247–276. Seattle, WA: University of Washington Press.

Pella, J., Masuda, M., Nelson, S. 1996. *Search algorithms for computing stock composition of a mixture from traits of individuals by maximum likelihood*. U.S. Department of Commerce, NOAA Tech. Memo, NMFS-AFSC-61, 68p.

Pella, J., Masuda, M. 2001. Bayesian methods for analysis of stock mixtures from genetic

characters. *Fishery Bulletin* 99: 151–167.
Pella, J., Masuda, M. 2005. Classical discriminant analysis, classification of individuals, and source population composition of mixtures. In *Stock identification methods: applications in fishery science*. (S. Cadrin, K. Friedland, J. Waldman, eds), pp. 517–552. New York: Academic Press.
Potvin, C., Bernatchez, L. 2001. Lacustrine spatial distribution of landlocked Atlantic salmon populations assessed across generations by multilocus individual assignment and mixed-stock analysis. *Molecular Ecology* 10: 2375–2388.
Raftery, A.E., Lewis, S.M. 1996. Implementing MCMC. In *Markov chain Monte Carlo in practice*. (W.R. Gilks, S. Richardson, D.J. Spiegelhalter, eds), pp. 115–130. London: Chapman and Hall.
Rannala, B., Mountain, J.L. 1997. Detecting immigrants by using multilocus genotypes. *Proceedings of the National Academy of Sciences, USA* 94: 9197–9201.
Wood, C., McKinnell, S., Mulligan, T., Fournier, D. 1987. Stock identification with the maximum-likelihood mixture model: sensitivity analysis and applications to complex problems. *Canadian Journal of Fisheries and Aquatic Sciences* 44: 866–881.
Worlund, D., Fredin, R. 1962. Differentiation of stocks. In Symposium on pink salmon. (N. Wilimovsky, ed.), pp. 143–153. *H.R. MacMillan Lectures in Fish. Vancouver, BC: University of British Columbia.*

APPENDIX E. Aggregating stocks for harvest management and an improved genetic stock identification

APPENDIX E. Aggregating stocks for harvest management and an improved genetic stock identification

Kenneth I. Warheit
Washington Department of Fish and Wildlife, 600 Capitol Way N., Olympia, WA 98501

Shawn Narum
Columbia River Inter-Tribal Fish Commission, 729 NE Oregon, Suite 200, Portland, OR 97206

Kathryn Kostow
Oregon Department of Fish and Wildlife, 17330 SE Evelyn Street, Clackamas, OR 97015

INTRODUCTION

Harvest management of Chinook salmon requires information on temporal and spatial distributions, exploitation rates, escapements, spawner abundance, productivity, and basic biology of stocks. Optimally, these data should be available for all stocks (here defined as biological populations). However, many of these variables are difficult to quantify on single stocks, and nearly impossible for all stocks potentially encountered in a managed fishery. A comprehensive knowledge of all stocks is unnecessary if stocks can be aggregated into groups.

All fishery analyses using coded wire tags (CWTs) are based on aggregates. For example, TCT models base exploitation rates on cohort analysis of CWTs and apply these rates to indicator stock complexes. These complexes or management groups do not necessarily include closely related stocks, but stocks of similar geography, run-timing, and management activities. The application of these indicator stock aggregates is based on at least two assumptions: (1) stocks in the aggregate co-occur in the fishery, and (2) data from the indicator stocks in the aggregate (*e.g.* exploitation rates) are applicable to the other stocks in the aggregate. These two assumptions are difficult to test, and may be violated if aggregates are simply *ad hoc* collections of populations designed to address single issues.

Fishery managers need to a construct aggregates that are useful for a suite of fishery management issues, where the managers can assume confidently that stocks within the aggregate share common characteristics that subject the stocks to the same or similar exploitation rates, for example. Closely related stocks with similar biology (*e.g.* run-timing) should have similar smolt development and outmigration timing, growth and development patterns, and ocean distributions, and should be subjected to the same or similar fishery pressures. Consequently, aggregates based on biology and recency of common ancestry (*i.e.* genetic similarity) may be groups that are best suited to address the needs of fishery management (even if recent common ancestry is a function of broodstock sharing among geographically distant hatcheries).

The goal of this section is to examine ways in which stocks can be aggregated into management groups. In addition, we consider the consequences of different aggregation methods on our ability to use GSI techniques to estimate mixture proportions or to assign individuals from a mixed stock fishery to stock aggregates. By way of example, we use stocks within Puget Sound to illustrate the effects of three aggregation procedures:

1. CTC indicator stock complexes

2. Puget Sound TRT stock groups

3. Statistical networks, a statistical method developed here

Our intention is not to advocate any particular method, but to provide a starting point for discussion on the purpose, methods, and consequences of aggregating stocks into management groups.

GENETIC DATA

We limited our statistical analysis to 25 Chinook stocks from Puget Sound and the Strait of Juan de Fuca (Table E1; Figure E1). The GAPS coastwide Chinook microsatellite baseline (v. 2.1) provides the foundation for this analysis, with the addition of collections genotyped by WDFW in the past year. These new data will be included in the next version of the GAPS baseline. Table 1 lists the stocks included in this analysis and indicates how the data used here differ from those in GAPS v. 2.1.

CTC INDICATOR STOCK COMPLEXES

Different stock groupings appear in various CTC reports. TCCHINOOK7-1 (2007, Appendix C.3) records three Areas for indicator stocks in Puget Sound: North/Central Puget Sound, Hood Canal, and South Puget Sound. Within these Areas there are Annex stock groups, Annex indicator stocks, escapement indicator stocks, exploitation rate indicator stocks, and model stocks. Alternatively, TCCHINOOK6-1 (2006, Table 7-1, for example) lists four Stock Complexes from the Puget Sound Area to summarize changes in the impacts of AABM fisheries on exploitation rate indicator stocks from 1979 to 2004: Puget Sound Spring (PSSp), North Puget Sound Fall (NPSFall), South Puget Sound Fall (SPSFall), and Hood Canal (HC) Chinook salmon. Based on the descriptions of these stock complexes, we placed 23 of the 25 GAPS stocks into one of these four Stock Complexes (Table E1; Figure E2). The Dungeness and Elwha stocks were not indicated as Stock Complexes, but were added to the analysis and grouped into a Strait of Juan de Fuca aggregate.

TRT STOCK GROUPS

The Puget Sound Technical Recovery Team (TRT) considered 22 populations of Chinook in Puget Sound (Ruckelshaus *et al*. 2006). All populations of Chinook in Puget Sound are considered part of a single ESU, and Ruckelshaus *et al*. (2006) were concerned mostly with describing individual populations. However, in their multidimensional scaling analysis, Ruckelshaus *et al*. (2006, Figure 6) aggregate the Puget Sound stocks into six groups. Although the authors do not suggest that these groups represent historical entities or should be used for fishery management, the groups include clusters of genetically similar stocks (although these clusters were not tested for statistical coherence). These aggregates are used here as a second example of how stocks can be grouped for fishery management (Table E2, Figure E3).

LIKELIHOOD-BASED STATISTICAL NETWORK

As an alternative to the two aggregate procedures described above, we designed a method based on Rannala and Mountain (1997). For each sample (*i.e.* individual fish), we calculated the Rannala and Mountain (1997) probability that its multilocus genotype occurred in each of the 25 stocks (*i.e.* 25 probabilities for each sample, scaled so that the probabilities for a single individual summed to one). X_{ij} = probability for the genotype of individual *i* occurring in population *j*. For each population, we calculated mean probability for each of the 25 stocks. That is, for each population we calculated $\left\{ \frac{\sum_{i=1}^{N_j} X_{ij}}{N_j}, \frac{\sum_{i=1}^{N_{j+1}} X_{ij+1}}{N_{j+1}}, \cdots, \frac{\sum_{i=1}^{N_{j+24}} X_{ij+24}}{N_{j+24}} \right\}$, a vector representing the mean probabilities that a multilocus genotype from that stock may occur in each of the 25 stocks (including itself). To determine the statistical significance of each of these mean probabilities, we randomly shuffled the probabilities for each individual, with respect to the populations. Therefore, if the original vector of probabilities for individual *i* = 1, with respect to *j* = 1–25 populations was

$$\{X_{1,1}, X_{1,2}, X_{1,3}, X_{1,4}, X_{1,5}, X_{1,6}, X_{1,7}, X_{1,8}, X_{1,9}, X_{1,10}, X_{1,11}, X_{1,12}, X_{1,13}, X_{1,14}, X_{1,15}, X_{1,16}, X_{1,17}, X_{1,18}, X_{1,19}, X_{1,20}, X_{1,21}, X_{1,22}, X_{1,23}, X_{1,24}, X_{1,25}\}$$

the randomly shuffled vector of probabilities for individual *i* = 1 could be

$$\{X'_{1,6}, X'_{1,10}, X'_{1,12}, X'_{1,3}, X'_{1,22}, X'_{1,14}, X'_{1,13}, X'_{1,24}, X'_{1,11}, X'_{1,20}, X'_{1,18}, X'_{1,5}, X'_{1,25}, X'_{1,7}, X'_{1,15}, X'_{1,1}, X'_{1,17}, X'_{1,9}, X'_{1,8}, X'_{1,21}, X'_{1,16}, X'_{1,23}, X'_{1,4}, X'_{1,19}, X'_{1,2}\}$$

Once the probabilities for all individuals were randomly shuffled, we calculated a new vector of mean probabilities for each population, and repeated the randomization procedure 10,000 times, producing 10,000 mean probability vectors for each population. If the true mean probability was equal to or greater than the 95th percentile of the randomized probabilities, we considered the true mean probability to be significant. For example, if the true mean probability for the Upper Skagit versus the Upper Cascade was significant (*i.e.* greater than the 95th percentile of the randomized probabilities), for any given multilocus genotype from the Upper Skagit there is a significantly greater probability that it will also occur in the Upper Cascade than it would occur in any randomly chosen population from Puget Sound. This suggests that there is either current or historical gene flow between these two populations. [Since the mean probability for a multilocus genotype from the Upper Skagit occurring in Upper Cascade may not be the same as the mean probability for a multilocus genotype from the Upper Cascade occurring in Upper Skagit, this method may be useful to test for asymmetric gene flow].

We can discover with this method networks of shared multilocus genotypes among populations within a defined geographic area such as Puget Sound. We graphically represented these networks by joining with a line those pairwise populations with significant mean probabilities (Figure E4). The interconnected lines reveal two main clusters of stocks in Figure E4 connected to each other at two points: Lower Skagit – Samish rivers, and Snoqualmie – Nisqually rivers. The Nooksack River Spring and White River Spring populations are wholly independent, and the Elwha and Dungeness river populations are mutually connected. Although the connections between the two large clusters create some ambiguity, five stock groups in total can be

distinguished (Figure E5, Table E1). These five aggregates are similar to the TRT groups, with the only exception being that the Puget Sound Spring/Summer group established by this method, was split into the Snohomish and Puget Sound Spring/Summer groups in the TRT analysis.

RELATIONSHIPS BETWEEN PROPOSED MANAGEMENT GROUPS AND COMMON ANCETRY/GENETIC SIMILARITY

We constructed a phylogenetic hypothesis for the evolutionary relationships of the Puget Sound stocks using an allele-sharing matrix for the 13 microsatellite loci, and a neighbor-joining tree[3]. The tree was rooted by two Middle Fraser River stocks (Stuart River Fall and Upper Chilcotin River Spring), with the assumption that these two stocks were outgroups with respect to a monophyletic[4] Puget Sound group (Figure E6). Stocks were labeled with a color-code signifying its management group identity, for each of the three alternative aggregating procedures. The intent was to determine if any of the aggregating procedures produced mutually exclusive monophyletic groups. Genetically similar groups are not necessarily monophyletic, so management groups are not expected to be exclusively monophyletic. However, in this exercise the stocks within a monophyletic group likely have similar development, life histories, behaviors, and ocean distributions, and would therefore have similar probabilities of occurring in particular fisheries[5]. If true, management groups consisting of monophyletic groups of populations would have greater predictive power than polyphyletic or paraphyletic[6] management groupings for determining the effects of a fishery on a particular stock. Hence, the use of monophyletic stock groups would be superior in a harvest management program.

None of the three aggregating procedures produced mutually exclusive monophyletic groups (Figure E6), but the Statistical Networks (SN) procedure was superior to the other two procedures. With both the TRT and SN procedures, the Puget Sound Spring/Summer group is paraphyletic with respect to the Skykomish, Snoqualmie, and Lower Skagit Fall runs. The TRT procedure also produced a paraphyletic Snohomish River group. All groups in the CTC procedure are paraphyletic except Hood Canal and Strait of Juan de Fuca, the latter of which was not considered by the CTC (Figure E6).

[3] Microsatellites are arguably not the best marker to reconstruct the phylogenetic history of salmonid stocks. Perhaps a better procedure would be to construct phylogenetic hypotheses using a suite of different marker and marker types, and use microsatellites (or SNPs), for example, to estimating aggregate proportions.

[4] A monophyletic group includes a common ancestor and all its descendents.

[5] The main point here is that stocks within aggregates based on monophyly are *assumed* to be encountered in the same fisheries. Although in this exercise we are limiting our discussion to the process of aggregating Chinook stocks, the methods may be applicable to any species of salmonids, even those species with highly divergent life histories within monophyletic groups (e.g., *Oncorhynchus mykiss*, *O. nerka*). For *O. mykiss* and *O. nerka*, for example, only one of the divergent life histories is subjected to ocean fisheries (steelhead and sockeye, respectively). Therefore, steelhead stock aggregates based on monophyly are indeed assumed to be encountered in the same fisheries.

[6] A paraphyletic group includes a common ancestor and some, but not all, its descendents, while a polyphyletic group includes taxa (*e.g.* stocks), but no common ancestor.

GENETIC STOCK IDENTIFICATION ERROR RATES

Genetic stock identification (GSI) error rates associated with each management group for the three aggregating procedures were estimated using the Anderson et al. (unpublished) CV-ML procedure. Here, 100% simulated mixtures were used with sizes for each stock that were 50% of the sample sizes listed in Table E1. For example, a 100% simulated mixture for North Fork Nooksack was N = 70. For each stock-based 100% simulated mixture, we estimated the proportion of the correct management group (e.g., for the Upper Cascade simulated mixture we estimated the Puget Sound Spring/Summer proportion in that mixture), and pooled all stocks within that management group to produce a single estimated proportion for the correct management group (*e.g.* results from the 100% simulated mixtures for all stocks within the Puget Sound Spring/Summer were pooled together to produce a single Puget Sound Spring/Summer estimate). This process was repeated 10,000 times to produce a distribution of estimated proportions (Figure E7).

As with the phylogenetic analysis in Figure E6, the SN performed best, producing the lowest error rates, while the error rates for the CTC model were the highest. The median value for four of the five management groups in the SN procedure was 1.00, with the value for the fifth group being 0.98. That is, for each group one-half of the 10,000 runs produced an error rate of 2% or less. The highest error rate for the analysis was 58% for the Hood Canal management group under the CTC procedure. The SN produced relatively good results, but several outlier runs appeared for each of the management groups, resulting in skewed distributions (note the relative positions of the median and mean values in Figure E7). The cause of these outlier runs is best illustrated with the Nooksack Spring group (Figure E8). Although the mixture proportion for 6,689 of the 10,000 simulated mixtures was correctly identified as 99% or greater (*i.e.* error rate of 1% or less), 1,102 of the runs produced stock proportions for the Nooksack Spring group as 10% or less (*i.e.* error rate of 90% or greater; see also Figure E7). The Nooksack Spring collection in the GAPS database (v. 2.1) apparently contains a higher than expected number of fall fish of Samish origin. Since each run of the simulation effectively resamples the baseline, the 1,102 runs that produced poor results contained a high proportion of these fall fish, resulting in a bimodal distribution (Figure E8).

CONCLUSION

Fishery managers use stock composition estimates to assess catch allocation compliance and harvest impacts. For the most part compliance and impacts are measured on aggregates of stocks rather than on specific stocks, unless ESA issues are a factor. As the CTC documents cited above demonstrate, there are a variety of reasons to aggregate stocks, and different stock aggregates are used to make different calculations pertaining to a fishery. However, the method used to aggregate stocks will affect the efficacy of genetic stock identification (GSI). That is, aggregation schemes that are inconsistent with the genetic relationships of the stocks will reduce the accuracy and precision of GSI, thereby limiting the usefulness of genetic analyses, and compromising our ability to manage fisheries with a full suite of data.

RECOMMENDATIONS

Standard quantitative stock aggregations should be designed *coast-wide* to accomplish two goals:

1. To be consistent with the phylogenetic relationships of stocks, and
2. To maximize value to address specific fishery management needs.

Table E1. List of stocks used in this analysis. The numbers correspond to location in Figure E1. Timing is the run timing for the stock: Spring (Sp), Summer (Su), and Fall (F). Origin refers to source of samples, either hatchery (H), or in-river (W).

	Stock Name	Timing	Origin	N	New Data[2]	Stock Grouping[3]		
						CTC	TRT	SN
1	North Fork Nooksack	Sp	HW	139		PSSp	Nooksack Sp.	Nooksack Sp.
2	Samish	F	H	82		NPSFall	PSSpSu	PSSpSu
3	Lower Skagit	F	W	108	1	NPSFall	PSSpSu	PSSpSu
4	Upper Skagit[1]	Su	HW	226	2	NPSFall	PSSpSu	PSSpSu
5	Upper Cascade	Sp	W	48	2	PSSp	PSSpSu	PSSpSu
6	Marblemount Hatchery	Sp	H	121	2	PSSp	PSSpSu	PSSpSu
7	Lower Sauk	Su	W	30		NPSFall	Snoh	PSSpSu
8	Upper Sauk	Sp	W	164	2	PSSp	PSSpSu	PSSpSu
9	Suiattle	Sp	W	152		PSSp	PSSpSu	PSSpSu
10	North Fork Stillaguamish	Su	HW	345		NPSFall	PSSpSu	PSSpSu
11	Skykomish	Su	HW	309	2	NPSFall	Snoh	PSSpSu
12	Snoqualmie	Su	W	54		NPSFall	Snoh	PSSpSu
13	Elwha	Sp	HW	388		SJF	SJF	SJF
14	Dungeness	Sp	W	132		SJF	SJF	SJF
15	Hood Canal (Hamma Hamma)	F	W	140		HC	PSFall	PSFall
16	Skokomish	F	HW	329	2	HC	PSFall	PSFall
17	Grover's Creek Hatchery	F	H	95	1	SPSFall	PSFall	PSFall
18	North Lake Washington (Bear Creek)	F	HW	237	1	SPSFall	PSFall	PSFall
19	Portage Bay (UW) Hatchery	F	H	140	1	SPSFall	PSFall	PSFall
20	Issaquah Creek	F	HW	229	1	SPSFall	PSFall	PSFall
21	Cedar River	F	HW	221	1	SPSFall	PSFall	PSFall
22	Green River (Soos Creek) Hatchery	F	H	184		SPSFall	PSFall	PSFall
23	Puyallup River	F	HW	198		SPSFall	PSFall	PSFall
24	Nisqually River	F	HW	238	2	SPSFall	PSFall	PSFall
25	White River (Puyallup)	Sp	HW	242		PSSp	White Sp.	White Sp.

[1] Includes Marblemount summer broodstock and natural spawning fish in the Upper Skagit River.

[2] New data indicates how data used in this analysis differs from that in GAPS v. 2.1. 1 = populations not included in GAPS 2.1. 2 = population included in GAPS v. 2.1, but additional samples from new collections added to GAPS 2.1 data.

[3] See text and Figures 2 (CTC), 3 (TRT), and 5 (SN).

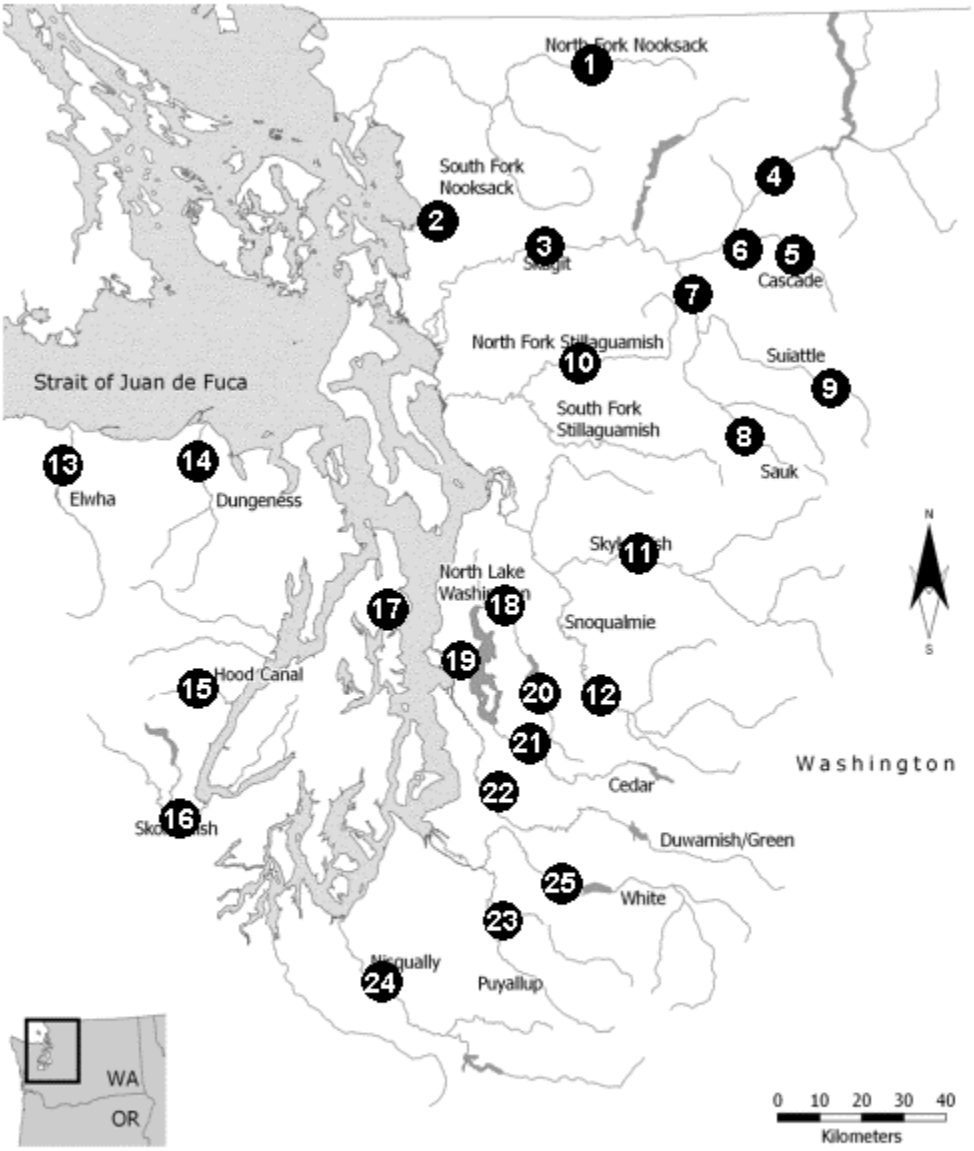

Figure E1. General location of stocks used in this analysis. See Table E1 for names of and additional information for each stock. Base map from Ruckelshaus *et al*. (2006).

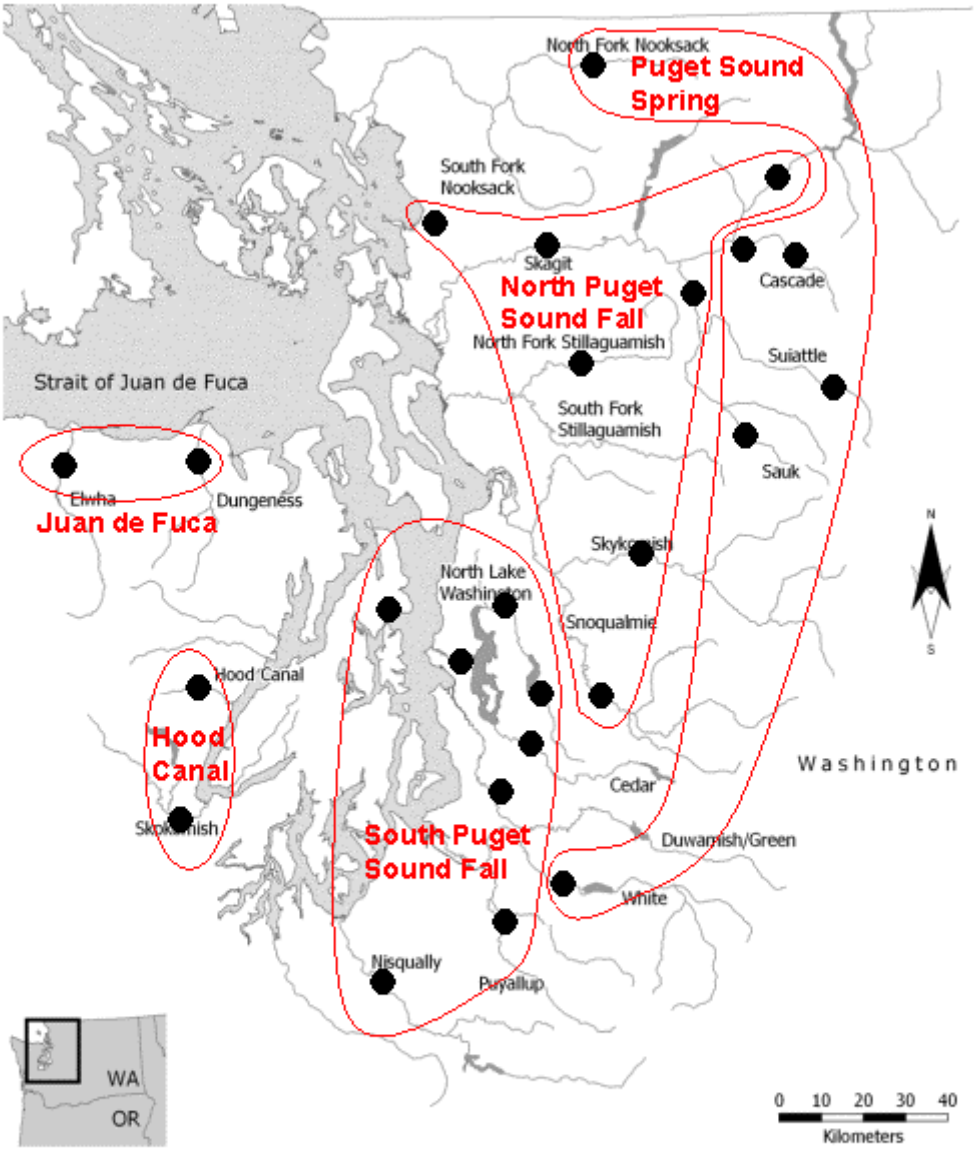

Figure E2. Stock aggregations, based on the CTC stock complex definitions. The Strait of Juan de Fuca group was not listed as a Stock Complex by the CTC, but added to this analysis. See Figure E1 and Table E1 for individual stock locations. Base map from Ruckelshaus *et al*. (2006).

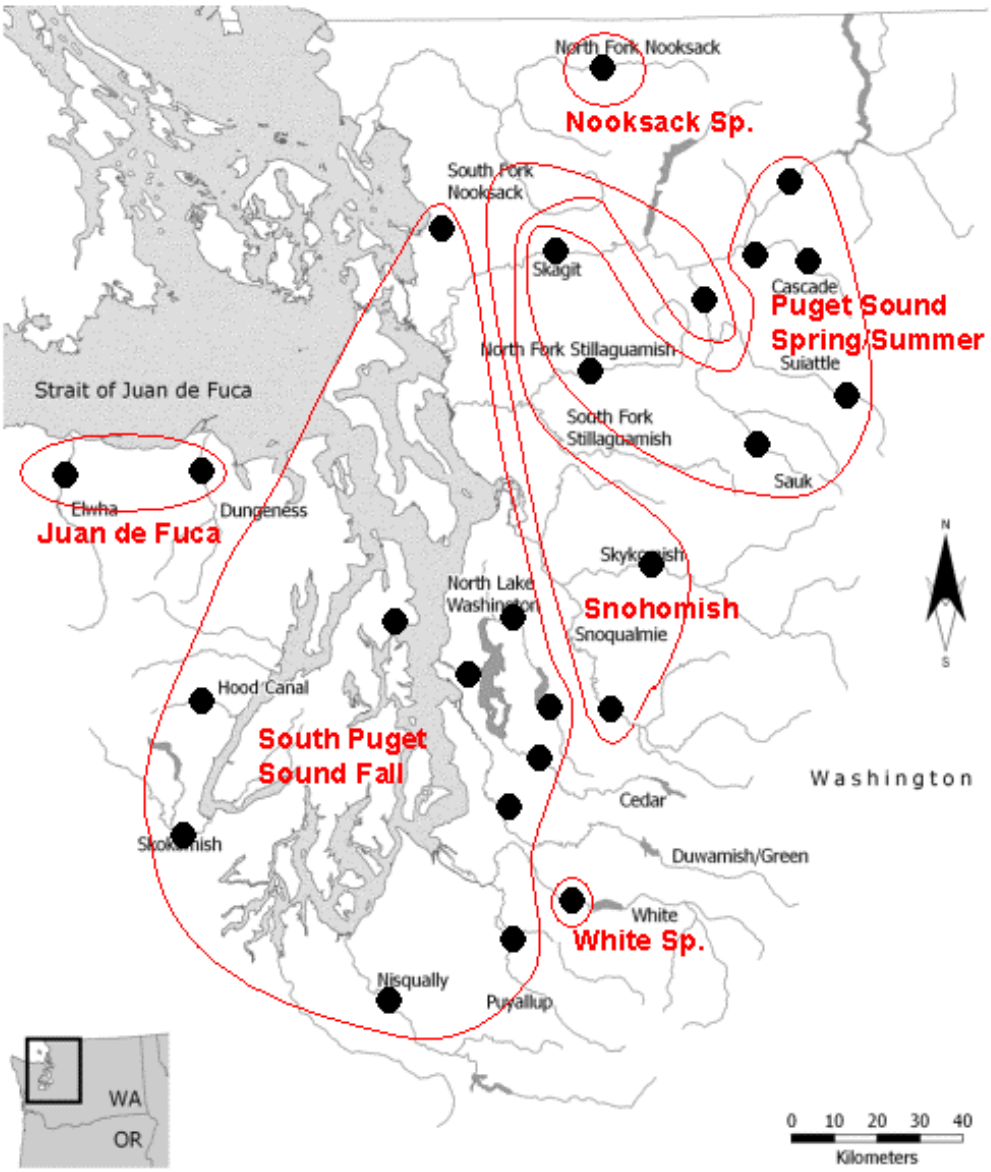

Figure E3. Stock aggregations, based on the TRT multidimensional scaling (see text). See Figure E1 and Table E1 for individual stock locations. Base map from Ruckelshaus *et al.* (2006).

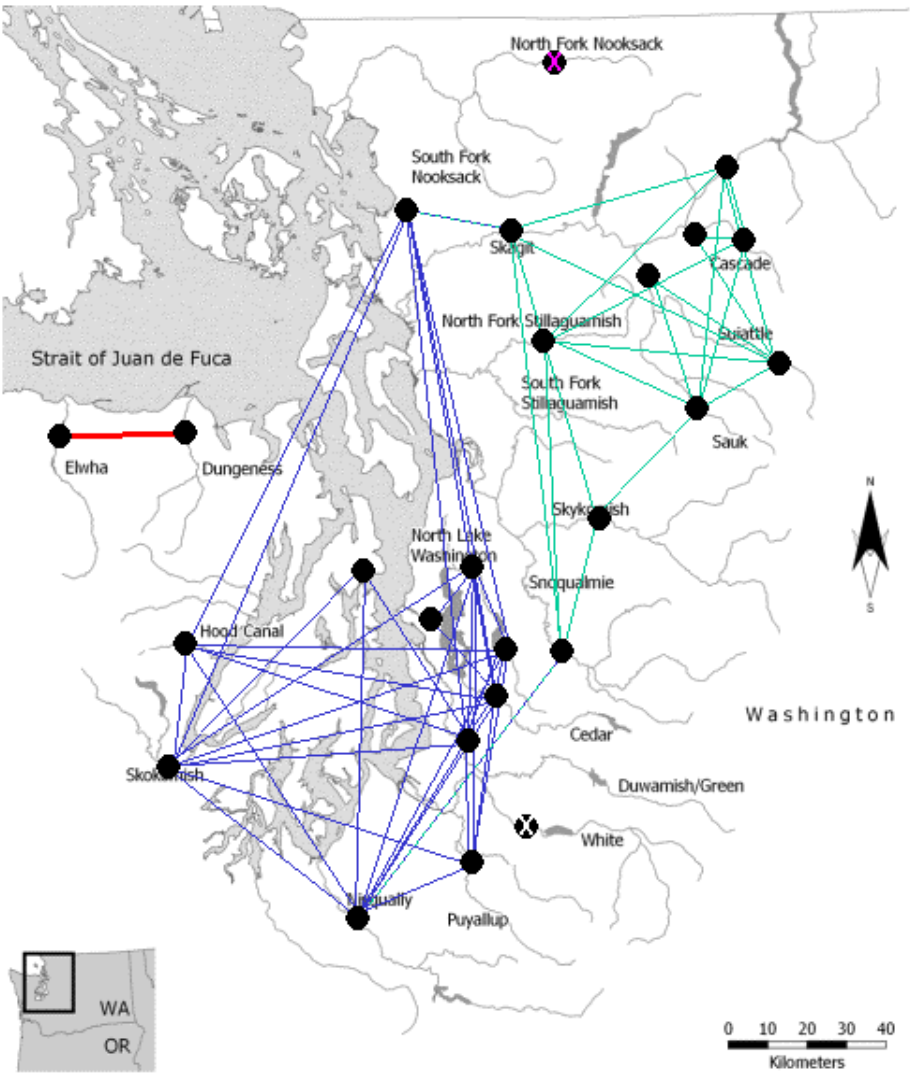

Figure E4. Results from SN procedure described in the text. Lines between pairs of stocks indicate mean probabilities significantly greater than random. That is, if a line connects two stocks, the mean of the probabilities of individuals from one stock (or both stocks) assign to the other stock is greater than expected from a random distribution of probabilities. Note, White and Nooksack Rivers without connecting lines. Individual networks shown in different colors (see also Figure E6). Base map from Ruckelshaus *et al*. (2006).

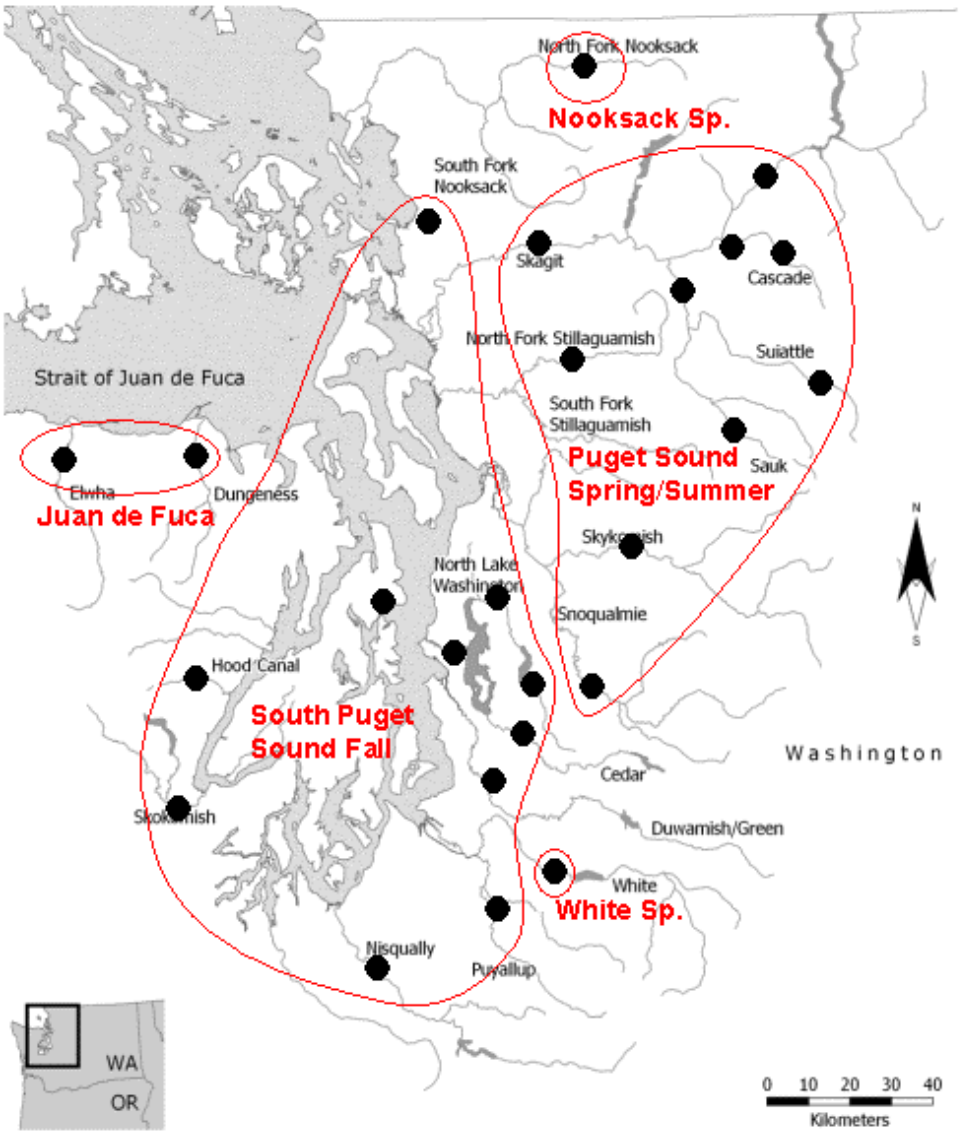

Figure E5. Stock aggregations, based on the Statistical Networks model described in the text, and shown graphically in Figure E4. Figure E1 and Table E1 for individual stock locations. Base map from Ruckelshaus *et al*. (2006).

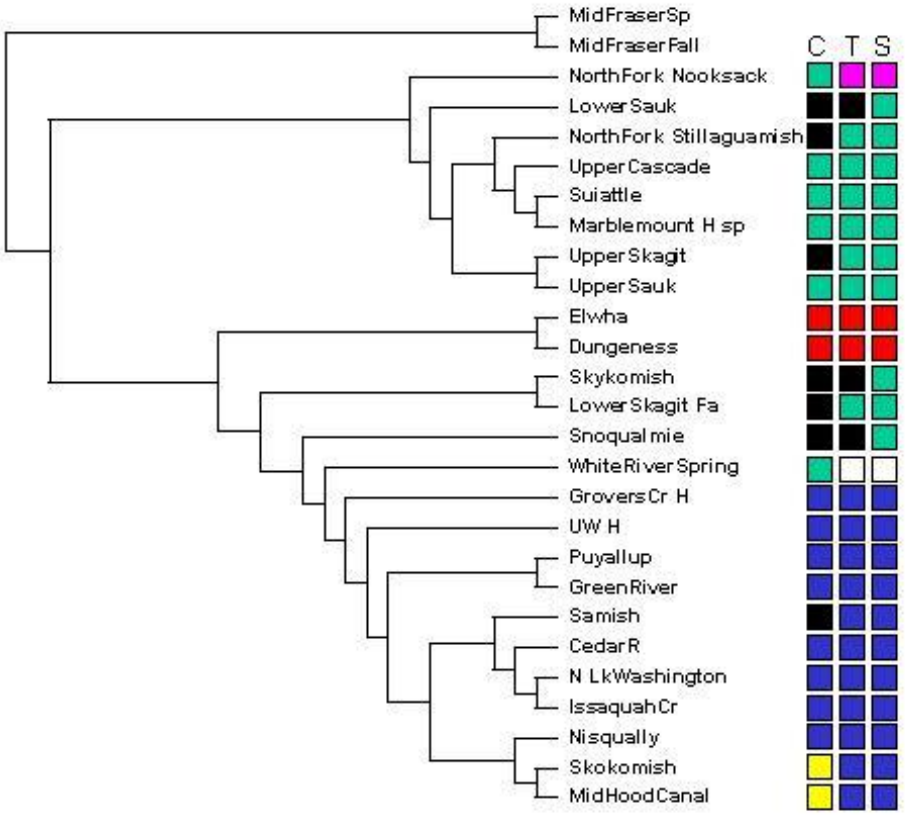

Figure E6. Neighbor-joining tree from an allele-sharing matrix, with two Fraser River stocks included as an outgroup. Colored-filled boxes are management group identities for each of the three alternative aggregating procedures (C = CTC, T = TRT, S = SN).

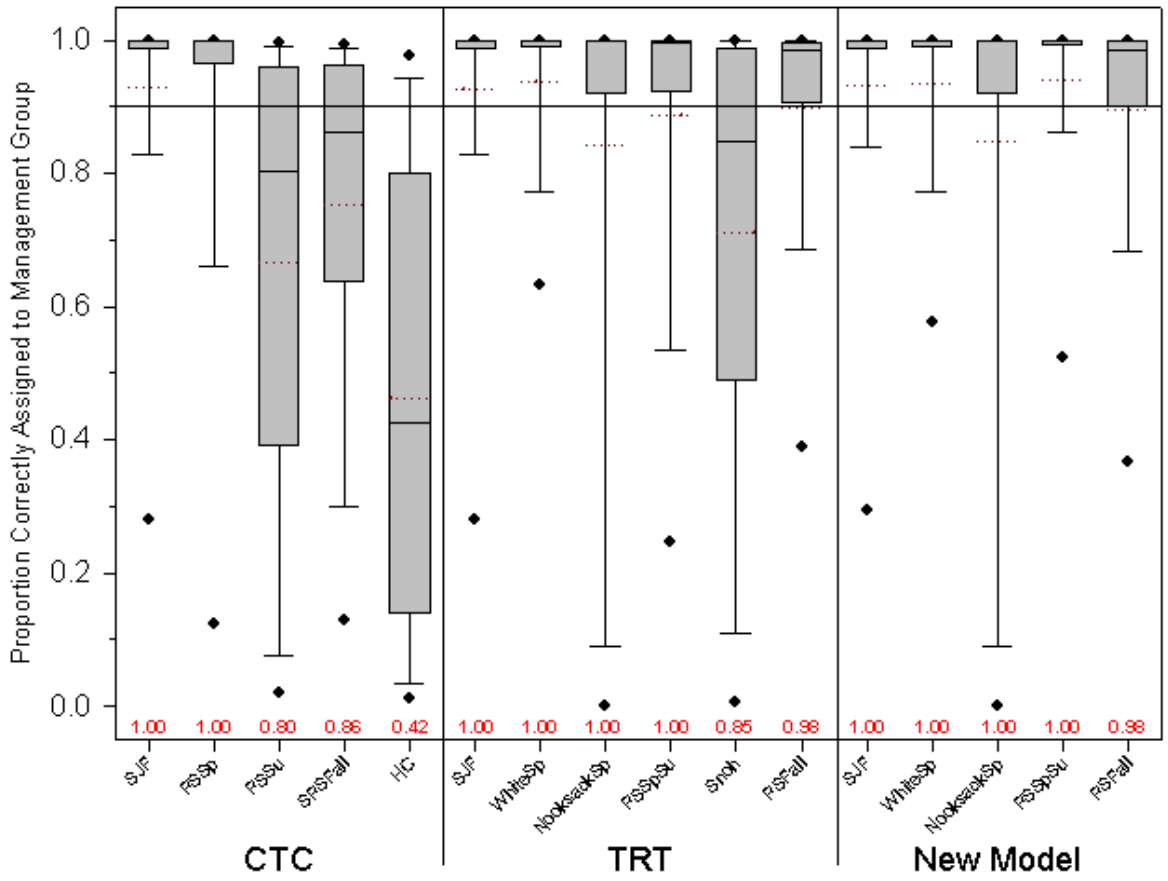

Figure E7. Box plots showing the distribution of correct assignments for 10,000 simulated 100% mixtures, for each management group. The box extends from the 25th to the 75th percentile, the bars cover the 10th and 90th percentile, and black dots are the 5th and 95th percentile for the 10,000 runs for each management group. The solid and dotted lines associated with each plot are the median and mean values, respectively, for the 10,000 runs. The median values for each management group are also written at the bottom of the plot above the group identification. The SJF, WhiteSp, NooksackSp, and PSFall groups for the TRT and New Method are identical, and therefore produced the same box plot.

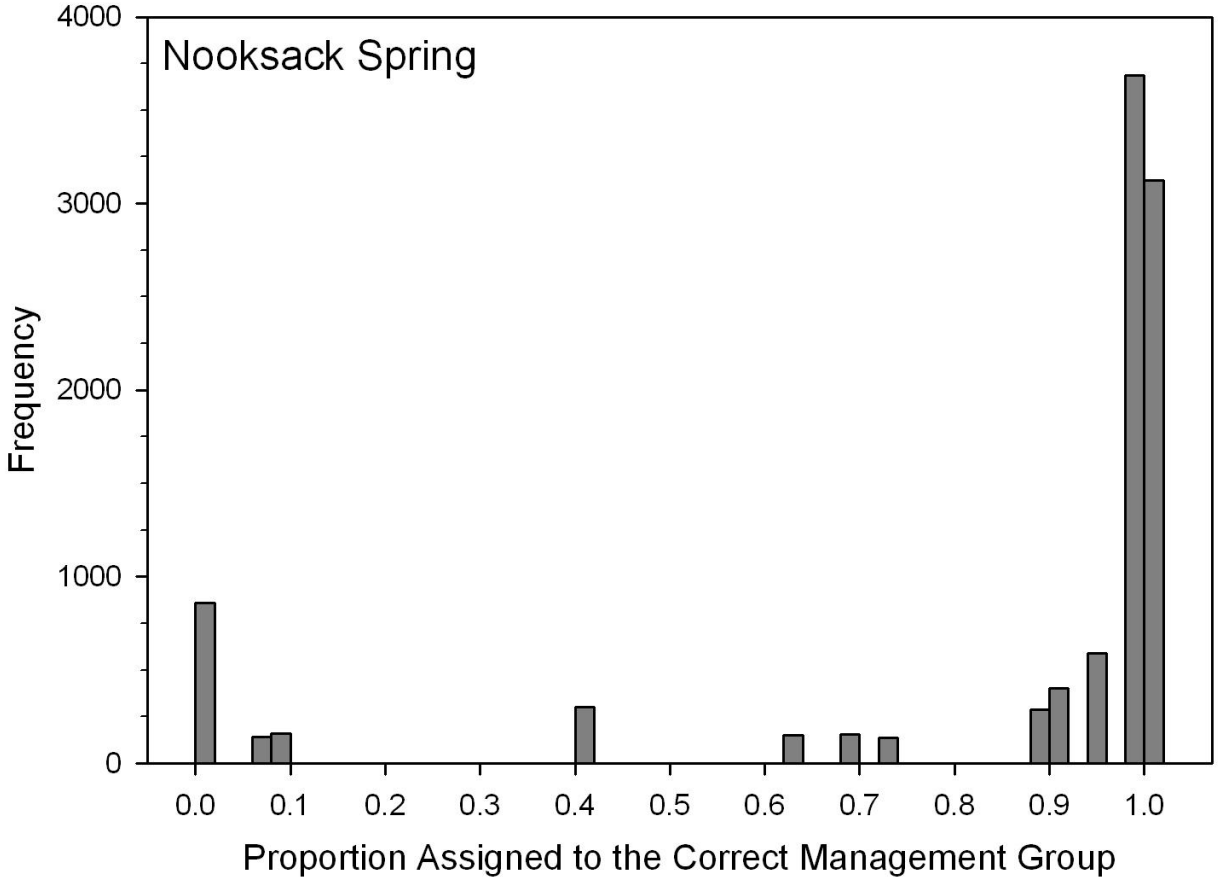

Figure E8. Frequency distribution for proportion Nooksack Spring samples correctly assigned to the Nooksack Spring management group for 10,000, 100% simulated mixtures. This plot shows the frequency distribution for the Nooksack Spring box plots in Figure E7.

CITATIONS

Rannala, B., Mountain, J.L. 1997. Detecting immigrants by using multilocus genotypes. *Proceedings of the National Academy of Sciences, USA* 94: 9197–9201.

Ruckelshaus, M.H., Currens, K.P., Graeber, W.H., Fuerstenberg, R.R., Rawson, K., Sands, N.J., Scott, J.B. 2006. Independent populations of Chinook salmon in Puget Sound. U.S. Department of Commerce, NOAA Technical Memorandum NMFS-NWFSC-78, 125 p.

APPENDIX E. HOW DIFFERENT SOURCES OF ERROR AFFECT THE ACCURACY OF GENETIC STOCK IDENTIFICATION

APPENDIX E. How Different Sources of Error Affect the Accuracy of Genetic Stock Identification

Steven Kalinowski

Department of Ecology, Montana State University, PO Box 173460, Bozeman, MT 59717-3460

INTRODUCTION

One important step in improving GSI accuracy is to identify sources of error affecting GSI estimates. Estimates can differ substantially from actual proportions in a mixture, as in 100% simulations of baseline data. Is this error due to sampling too few fish, to not genotyping enough loci, or perhaps to an inadequate baseline of contributing populations? Answers to these questions would help to improve GSI results. A recently development method is used here to decompose the total expected square error into important sources of error.

SOURCES OF GSI ERROR

GSI estimation error can arise from several factors (Table F1). One source of error occurs during the sampling of the fishery. Sampling may be biased toward fish from some populations, but not others. This error can be reduced by enlarging samples sizes or by sampling over larger periods of time or from a greater number of fishing boats. A second source of error can occur even when the fishery is sampled randomly; the random sampling simply fails to include all stocks present in the fishery. Larger fishery sample sizes may reduce this source of error. A third source may be due to the sampling of a finite number of loci. The inclusion of additional markers can potentially help to reduce this error. A fourth source of error can arise from genotyping error in the laboratory. A fifth source arises from errors on allele frequencies that arise from sampling a finite number of individuals in baseline populations. The last source of error is the presence of fish in the fishery sample from population that have not been included population in the baseline. Here, three of these sources of error, fishery sample size (2), locus sampling (3), and baseline allele-frequency estimation (5) are examined in a set of empirical data for a Chinook salmon fishery off southeastern Alaska.

Table F1. Potential sources of GSI error.

a. Fishery sampling bias
b. Finite sampling of fishery
c. Finite sampling of genetic markers
d. Genotyping error
e. Finite sampling of baseline populations

ERROR DECOMPOSITON[7]

A convenient measure of how much estimates are expected to be wrong is total expected square error (ESE)

$$ESE(\hat{\theta}_i) = E\left[(\theta_i - \hat{\theta}_i)^2\right]$$

This is like a variance, but also includes the effect of bias. The goal is to partition ESE into three components

$$ESE(\hat{\theta}_i)_{total} = ESE(\hat{\theta}_i)_{fishery} + ESE(\hat{\theta}_i)_{genotypic} + ESE(\hat{\theta}_i)_{baseline}$$

Total ESE is estimated with simulations using a 'conventional' method assuming allele frequencies in the baseline are known with certainty. Estimates of these three sources of error can be used to improve GSI accuracy. For example if $ESE(\theta_i)_{baseline}$ is small, then increasing sample sizes of baseline populations will add little to GSI accuracy. Likewise if $ESE(\theta_i)_{fishery}$ or $ESE(\theta_i)_{genotypic}$ are relatively large then expanding fishery sampling or the number of genetic markers can reduce GSI error.

Using this approach, total error was decomposed for the Chinook fishery in SE Alaska. First total ESE can be calculated with 'conventional' simulation methods assuming alleles frequencies are known

$$ESE(\hat{\theta}_i)_{total} \widehat{=} \frac{1}{R}\sum\left[(\theta_i - \hat{\theta}_i)^2\right]$$

$ESE_{fishery}$ due to fishery sampling can be calculated from the binomial variance based on sample size (N).

$$ESE(\hat{\theta}_i)_{fishery} = \frac{\theta_i(1-\theta_i)}{N}$$

Calculating the portion of the ESE due to baseline deficiencies requires knowledge of baseline allele frequencies. However, the problem is that these allele frequencies are unknown. One way around this problem is to adjust the population allele frequencies (GAPs frequencies in this case) to account for the increase in apparent divergence among populations due to finite sampling. A recently developed formula that adjusts allele frequencies towards the mean to reduce variance appears to work well (Kalinowski, unpublished).

[7] Software for the methods outlined here will soon be available on the Web at: www.montana.edu/kalinowski.

$$\begin{cases} \text{If } \hat{p}_i > \bar{p}, \quad \tilde{p}_i = \bar{p} + \sqrt{(\hat{p}_i - \bar{p})^2 - \left(1 - \frac{1}{k}\right)\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}} \\ \text{If } \hat{p}_i < \bar{p}, \quad \tilde{p}_i = \bar{p} - \sqrt{(\hat{p}_i - \bar{p})^2 - \left(1 - \frac{1}{k}\right)\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}} \end{cases}$$

These new frequency estimates are subtracted from total ESE to yield ESE due to baseline sampling

$$ESE\left(\hat{\theta}_i\right)_{baseline} = ESE\left(\hat{\theta}_i\right)_{total} - ESE\left(\hat{\theta}_i^*\right)$$

$ESE_{genotypic}$ is obtained by subtraction, since the three variance components are additive.

Table F1. Decomposition of GSI error in fishery samples of Chinook salmon from southeastern Alaska.

Source of error	Proportion
Fishery	9.5%
Genotypic	2.7%
Baseline	87.5%

RESULTS

The results of these simulations to estimate GSI error in a Chinook salmon fishery in southeastern Alaska appear in Table F2. Surprisingly, a large portion of the error is due to uncertainties in allele frequencies in baseline populations. A smaller proportion is due to fishery sampling and a very small proportion is due to genotypic sampling.

DISCUSSION

The results of this preliminary study show that errors from baseline sampling, and not fishery or genotypic sampling, are the major sources of error in these GSI estimates. In this particular case, accuracy would be improved only marginally by adding more loci or by enlarging the fishery sample. A greater improvement in accuracy can be achieved by improving baseline estimates. Preliminary results of other simulations appear to indicate that greater baseline sampling effort is justified when levels of divergence between contributing populations are small (*e.g.* $F_{ST} < 0.01$). In other circumstances, larger baseline samples will add little accuracy to GSI estimates. As GSI estimation in southeastern Alaska is typical of GSI estimation in several other fisheries, these results likely apply generally to other fisheries.

The effect of baseline sampling on the accuracy of GSI estimation has been investigated by several authors. For example, simulations show that the sampling of more than 50 individuals per population in a Columbia River steelhead baseline was unlikely to increase the accuracy of GSI estimates (Winans *et al*. 2004). This indicates that baseline sampling error is likely to be minimal for population sample sizes greater than 50. While baseline sample sizes of 100 appear to be sufficient for many GSI applications (Kalinowski 2004), recent work demonstrates the benefit of increasing baseline sample sizes when the level of genetic differentiation among populations is low (Kalinowski, unpublished). If baseline populations are similar (*e.g.* $F_{ST} < 0.005$), increasing baseline sample sizes to 200 or more may be needed to increase the accuracies of GSI estimates.

If baseline allele-frequency estimation is in fact the largest source of GSI error, the focus should be on ways of improving the accuracies of these estimates. Two ways might be used without sampling more fish.

1. A generalized expectation maximum algorithm would use mixture samples to improve estimated allele frequencies in the baseline population.

2. Spatial models of allele frequencies are based on the tendency of populations near to each other to be similar, and can be adapted from mathematical modeling used in epidemiology.

Other problems were not addressed with the simulations presented here. The most vexing source of error is unsampled source populations, and it is difficult to use simulations to estimate how unsampled populations impact GSI estimation. Nevertheless, three approaches could be used to evaluate the magnitude of this problem. First, simulations could be made to examine the impact of excluding some existing populations from baselines while keeping them in mixtures to which GSI is applied. Second, spatially explicit models of population structure could be constructed to estimate allele frequencies of unsampled populations, and these estimates could be used in conventional GSI simulations. Third, a Bayesian missing-data model may successfully identify unsampled populations contributing to a fishery (Pella and Masuda 2006).

Another problem not addressed here is error in the estimation of the frequencies of low-contributing stocks. This problem is challenging for two reasons. First, if a stock of fish is rare in a fishery, a large fishery sample size will be needed to accurately reflect the composition of the fishery. Second, the statistical approaches used to perform GSI tend to bias stock composition estimates toward $1/k$, where k is the number of stocks contributing to the baseline. Hence, estimates of the frequency of a rare stock in a fishery will be biased upward. This bias is greatest when a rare stock is genetically similar to an abundant stock, because fish from the abundant stock are likely to be "mistaken" for fish from the rare stock; this will not be balanced by mistakes in the other direction because there are fewer individuals of the rare stock to be misidentified.

Both simulations and empirical approaches can be useful, but recent experience has demonstrated the limitations of relying on simulation alone. When the neutrality of alleles is assumed in a simulation model, other factors influencing GSI estimates may be unrecognized. In particular, the choice of high-graded markers showing large differences among populations can violate assumptions of neutrality, as these markers may be under the influence of natural selection or may be neutral but show larger than average differences between populations. Hence, the examination of empirical datasets may provide the best means of assessing power and of identifying components of GSI error.

CITATIONS

Kalinowski, S.T. 2004. Genetic polymorphism and mixed stock fisheries analysis. *Canadian Journal of Fisheries and Aquatic Sciences* 61: 1075–1082.

Pella, J., Masuda, M. 2007. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences* 63: 576–596.

Winans, G.A., Paquin, M.M., Van Doornik, D.M., Baker, B.M., Thornton, P., Rawding, D., Marshall, A., Moran, P., Kalinowski, S.T. 2004. Genetic stock identification of steelhead in the Columbia River basin: an evaluation of different molecular markers. *North American Journal of Fisheries Management* 24: 672–685.

APPENDIX F. INTRA- AND INTER-ANNUAL VARIATION IN STOCK COMPOSTION OF THE QUEEN CHARLOTTE ISLAND TROLL FISHERY 2002–2006

APPENDIX F. INTRA- AND INTER-ANNUAL VARIATION IN STOCK COMPOSITION OF THE QUEEN CHARLOTTE ISLAND TROLL FISHERY 2002–2006

Terry Beacham
Pacific Biological Station, Department of Fisheries and Oceans, Nanaimo, BC, V9R 5K6

INTRODUCTION

Results from mixed-stock analysis provide an opportunity to understand important features of salmon migration and spawning biology. Several datasets are now available that can be used to make these inferences. The following is an in-depth examination of data for an area off northern British Columbia that has been studied for several years and offers excellent insights into the migrations of particular Chinook salmon populations through an area.

Stock compositions were examined for Chinook salmon caught in either test troll fisheries or commercial troll fisheries off the northwest coast of the Queen Charlotte Islands. Samples were pooled on a monthly basis from April through September 2002–2006, with the regional stock composition determined as the mean of each monthly estimate over the five-year interval. Regional stock compositions by month are plotted in a series of figures, and the monthly changes in relative contributions are also indicated for the major stocks. In 2005, samples were unavailable from April, so a late March sample was considered as indicative of an April sample. Samples were also unavailable for July 2003.

IN-SEASON VARIABILITY

Stock composition of the samples analyzed clearly varied over the monthly sampling cycle. On a relative scale, Chinook salmon from Oregon were least prevalent in April (6%) and the most prevalent in September (58%), with the proportion of Oregon Chinook salmon increasing in every month during sampling (Figures G1–G6, G7). Similarly, Chinook salmon from Washington were relatively least abundant during April (4%), but the proportion of Washington populations increased in every month sampled, with the maximum proportion in September (21%) (Figure G8, G9). Chinook salmon from California followed the same trend as those from Washington and Oregon, with higher proportions (0.2%) observed in the later sampling months.

Columbia River Chinook salmon comprised an important component of the fish sampled from April through August. The maximum relative abundance of Columbia River populations appeared in April (44%) and the least in September (11%), but they constituted 23–29% of the monthly sample from May through August (Figure G8).

Fraser River Chinook salmon displayed a trend in relative abundance different from that observed in Oregon and Washington populations. Fraser River Chinook salmon were relatively most prevalent in May, June, and July, including between 27–36% of the fish sampled in these months (Figure G10). The peak Fraser River composition was observed in June (36%). By September, only about 3% of the fish sampled were estimated to be of Fraser River origin. Chinook salmon from the east coast of Vancouver Island comprised a minor component of the fish sampled in any month, ranging from 4% in April to 1-2% in the other months. Chinook salmon from the west coast of Vancouver Island were most prevalent in April (19%) and least prevalent in September (3%), with June and July the next months of least prevalence (6–7%) (Figure G11).

Chinook salmon from northern British Columbia were most prevalent in April (7%), and declined sequentially in each month to a low of 1% in September (Figure G12). Skeena River Chinook salmon were most prevalent from April through June (4–6%), and subsequently comprised about 1% of the monthly samples. Nass River and transboundary Chinook salmon displayed a similar pattern, each with higher prevalence from April to June (1–3%), and subsequently < 1% of the monthly sample.

The analyses of samples from troll fisheries indicated that stock compositions of the sampled fish changed during the course of the season. Very different stock compositions were observed in April as compared with the September samples. Monthly changes in stock composition likely reflected the migration patterns of various stock groups past the Queen Charlotte Islands.

INTER-ANNUAL VARIABILITY

Inter-annual variation was illustrated for five major stocks sampled from either the troll or commercial catches. Inter-annual variation in estimated stock composition was most pronounced for Oregon Chinook salmon in August, with the estimated proportion of the Oregon component ranging from 15–45%, and with the highest proportions observed in the August 2004 samples (Figure G13). The proportion of Columbia River Chinook salmon in the monthly samples was reasonably stable, with the greatest variance observed in May and August (Figure G14). A similar degree of variation was observed for Washington Chinook salmon, with relatively high proportions for the 2002 return year in all months (Figure G15). Fraser River Chinook salmon displayed relatively high proportions in all months during 2002 and 2006, likely reflecting the strong returns to the Thompson River drainage in those years (Figure G16). The proportions of west coast Vancouver Island displayed some variation, but the variation in the April samples may have been accentuated by the inclusion of the March 2005 sample in the analysis (Figure G17). In most months, the proportion of WCVI Chinook salmon was small relative to other major stocks present, and thus the absolute level of annual variation for this stock was less in comparison the other stocks.

DISCUSSION

This detailed examination of in-season and inter-annual variation stock-compositions in the Queen Charlotte Island fisheries illustrates the potential for using GSI estimates to evaluate the abundances and migration patterns of Chinook salmon. The comparisons presented here were possible only after a large-scale population baseline was established so that all the stocks potentially contributing to a fishery could be identified. While proportion estimates are an important starting point, abundance data or additional sampling may be required to extrapolate the results of a comparison such as this to other regions or fisheries. For example, additional samples may be needed to account for the non-random sampling by the troll fishery, as populations may be possibly differentially exploited by troll gear. Abundance data are also required to refine inferences of distribution and migration patterns. An important result of in-season and inter-annual comparisons is that sporadic sampling during a fishing season may give an incomplete view of the presences of various stocks contributing to a fishery.

Figure G1. April: GSI estimates for Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

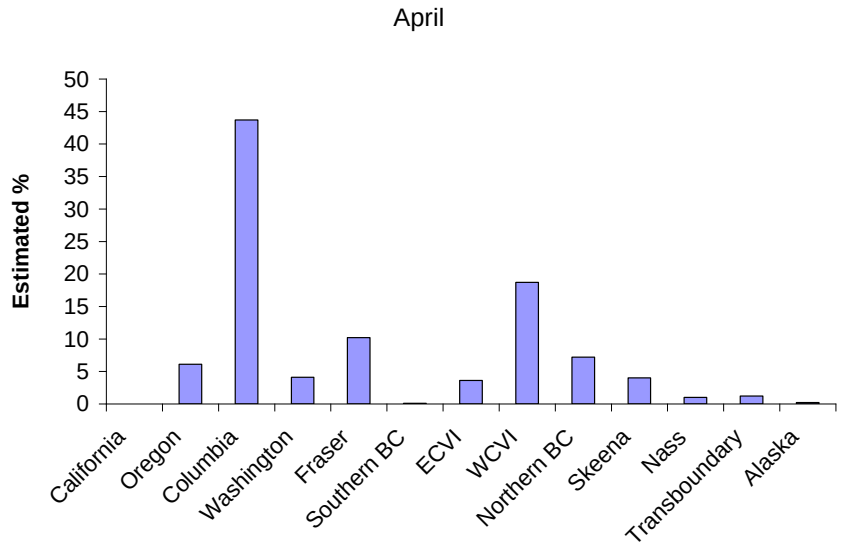

Figure G2. May: GSI estimates for Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

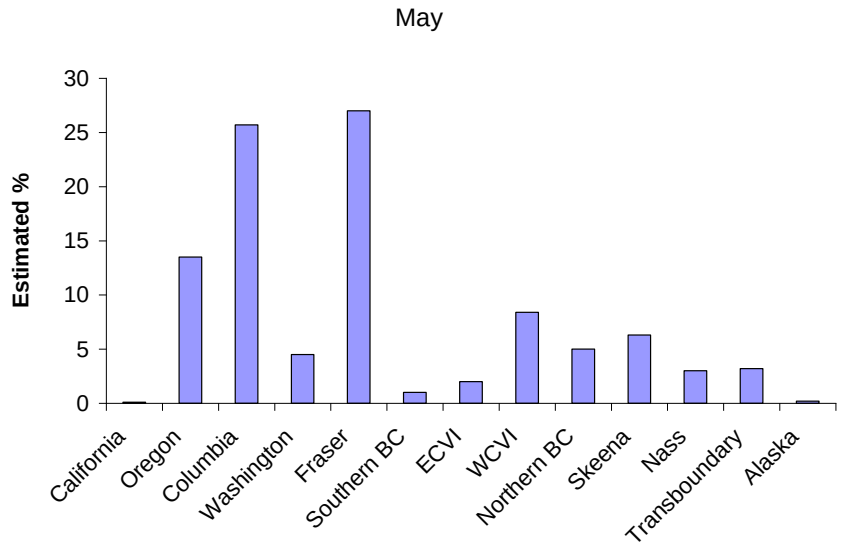

Figure G3. June: GSI estimates for Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

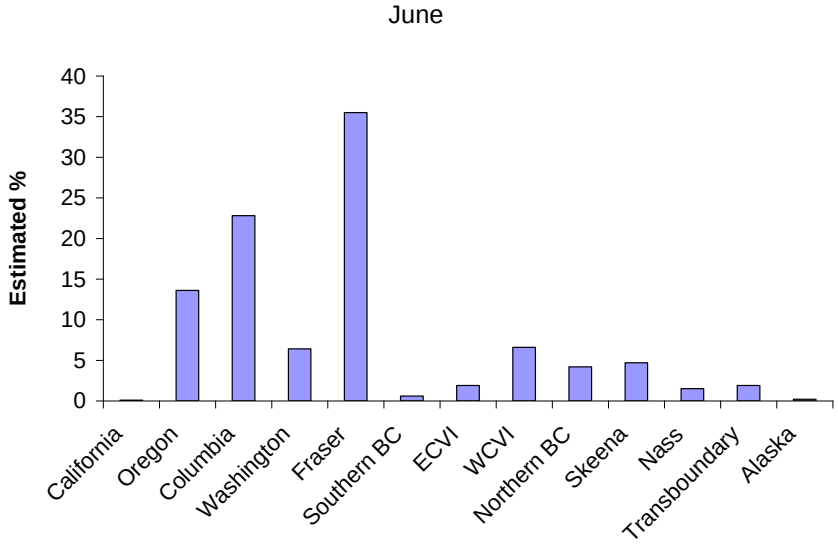

Figure G4. July: GSI estimates for Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

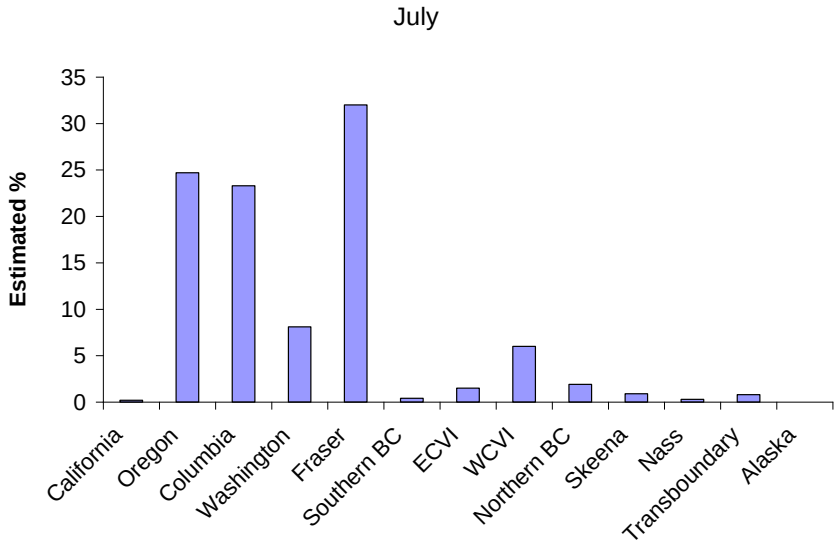

Figure G5. August: GSI estimates for Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

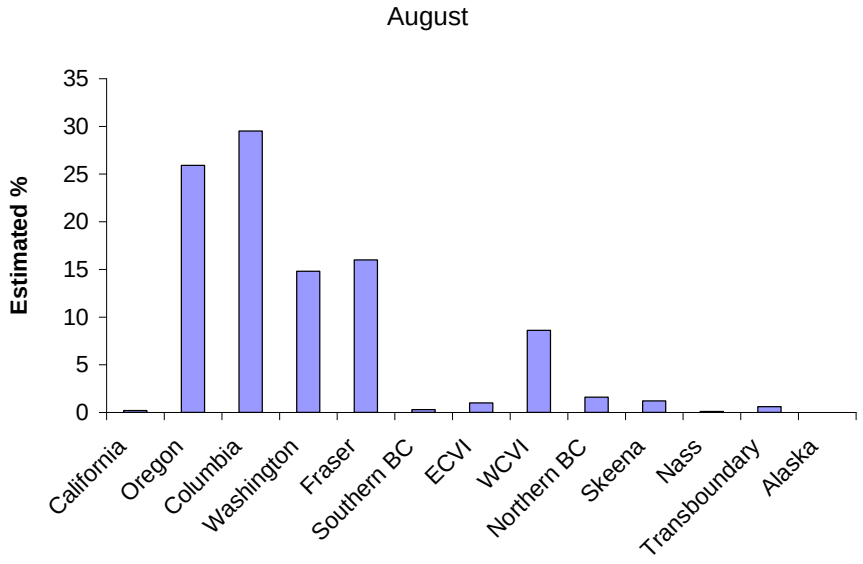

Figure G6. September: GSI estimates for Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

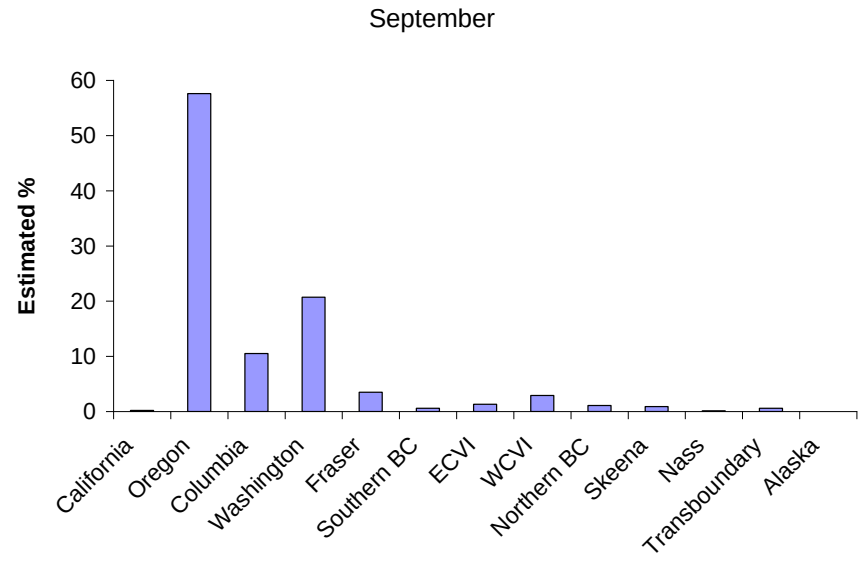

Figure G7. Summary of GSI estimates for fish from Oregon in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

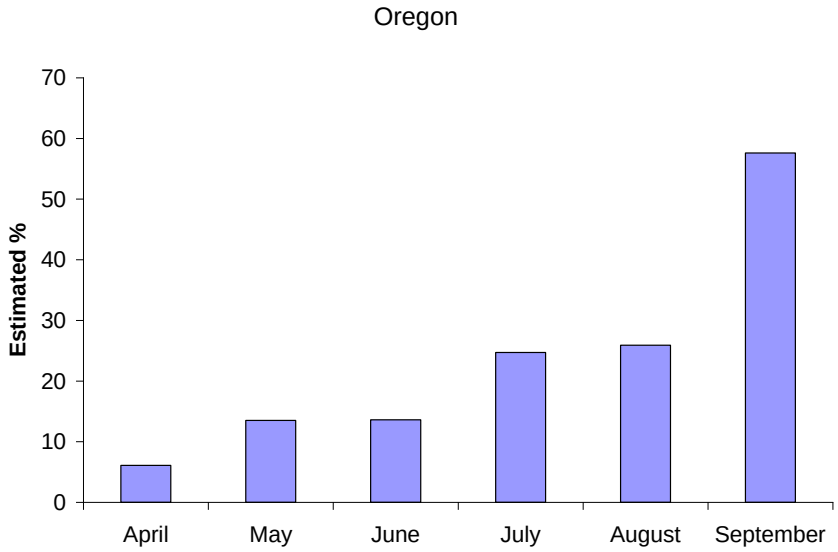

Figure G8. Summary of GSI estimates for fish from the Columbia River in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

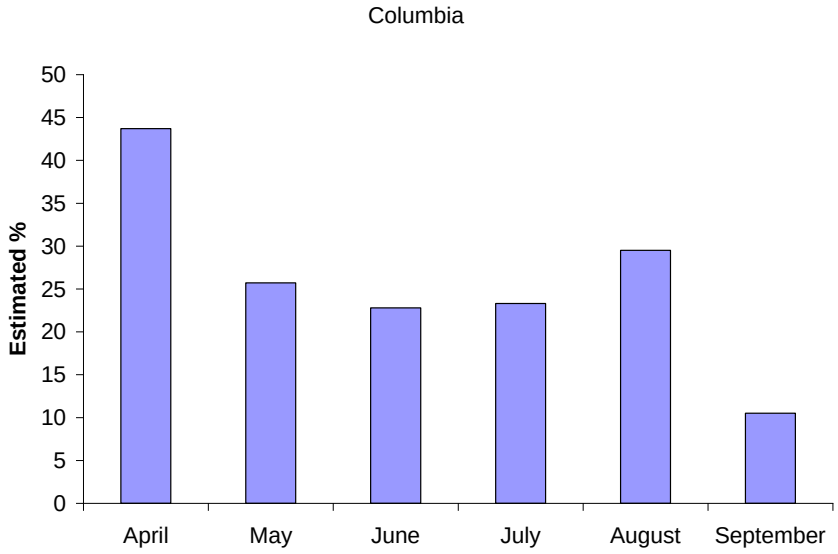

Figure G9. Summary of GSI estimates for fish from Washington State (non-Columbia River fish) in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

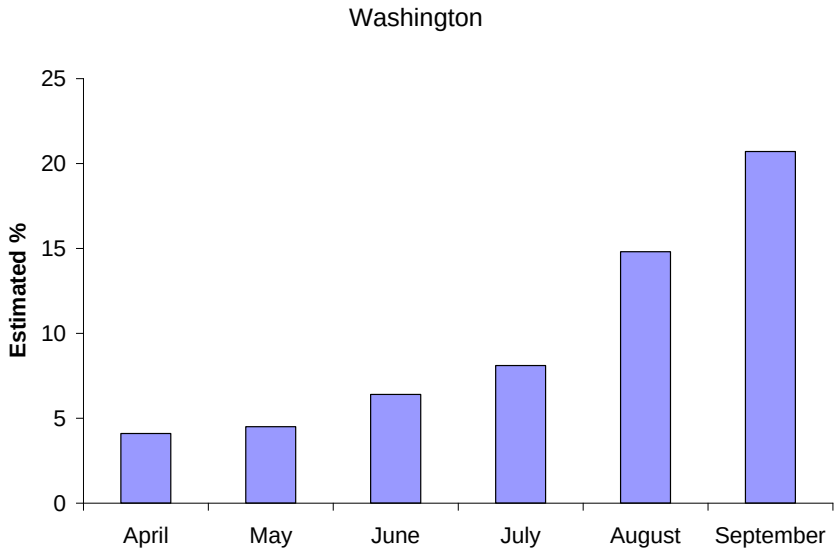

Figure G10. Summary of GSI estimates for fish from the Fraser River in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

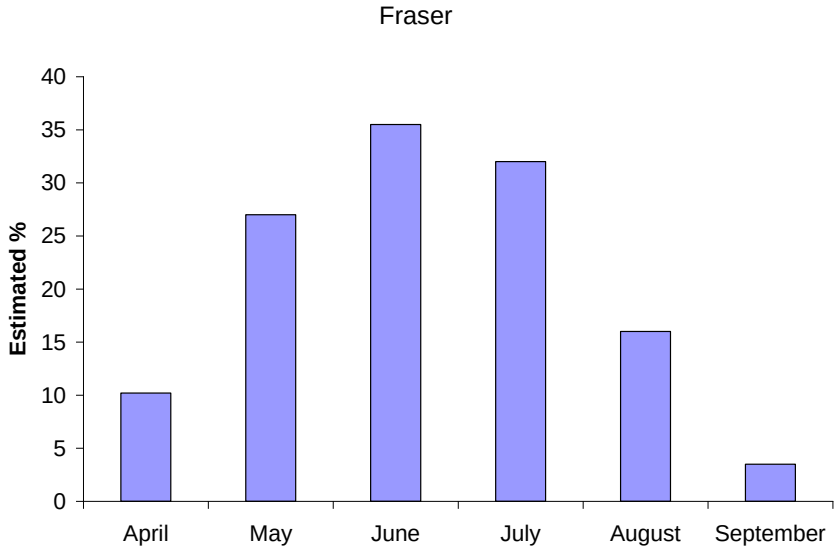

Figure G11. Summary of GSI estimates for fish from west coast of Vancouver Island (WCVI) drainages in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

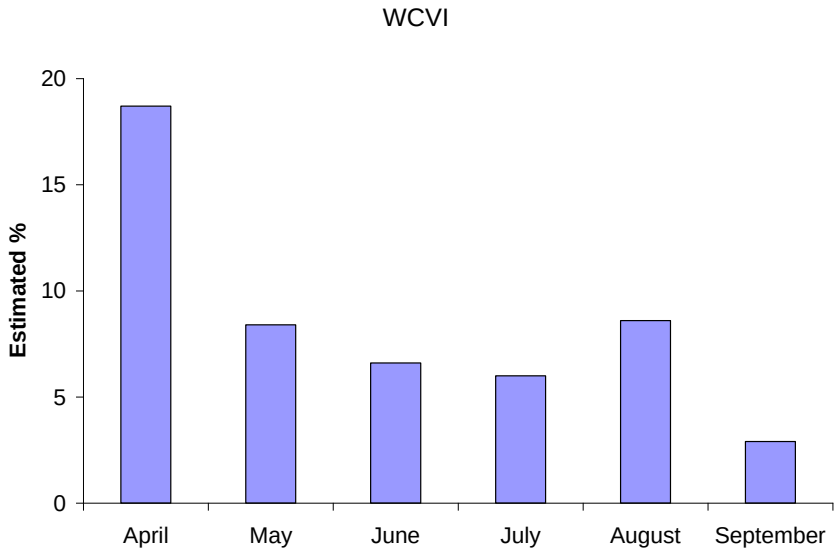

Figure G12. Summary of GSI estimates for fish from the northern British Columbia drainages in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

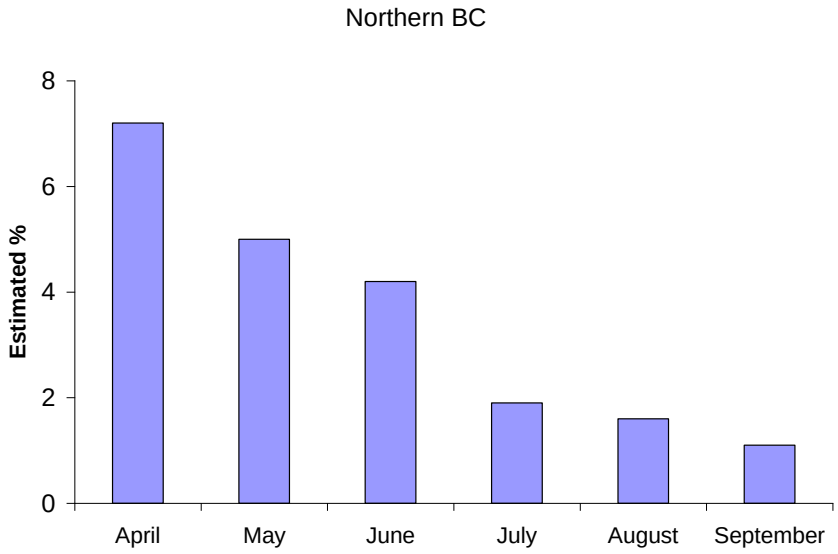

Figure G13. Inter-annual variability (2002–2006) in GSI estimates of fish from Oregon State in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

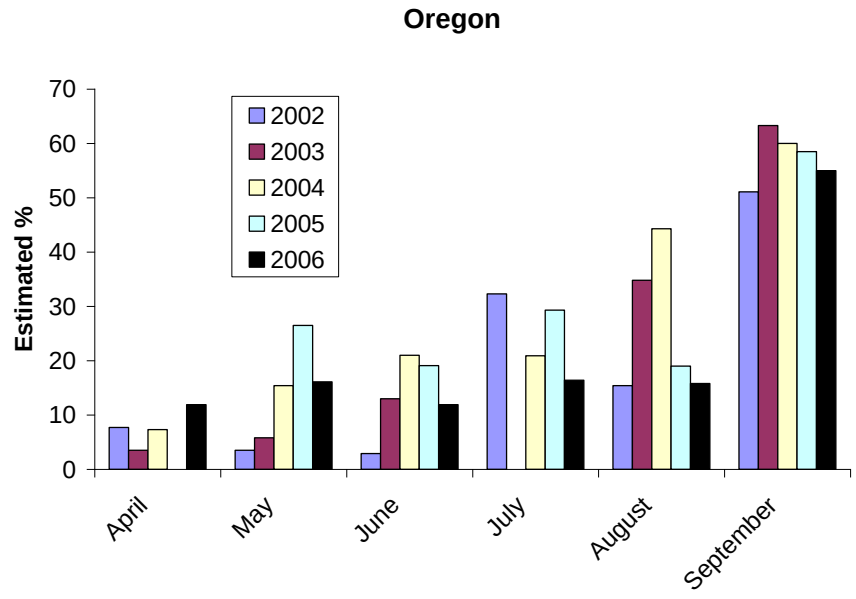

Figure G14. Inter-annual variability (2002–2006) in GSI estimates of fish from the Columbia River in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

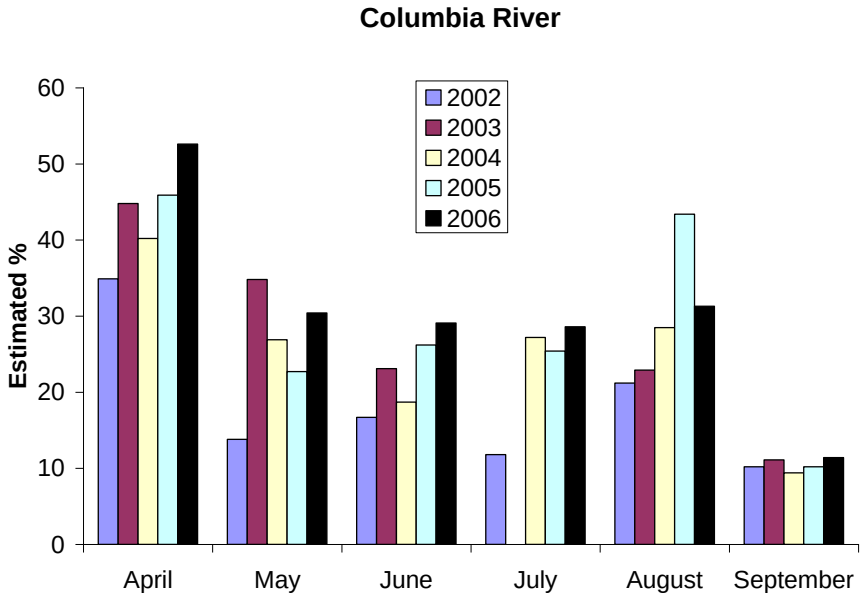

Figure G15. Inter-annual variability (2002–2006) in GSI estimates of fish from Washington State in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

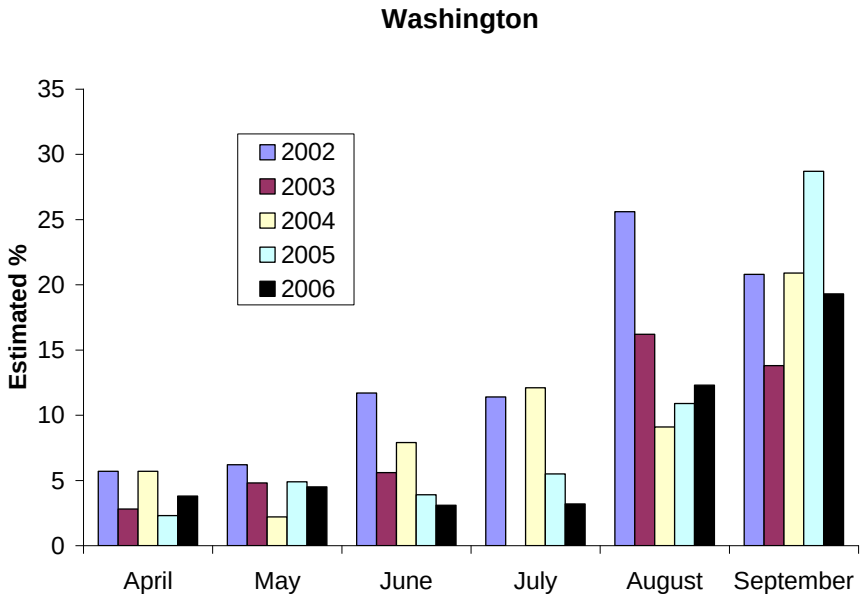

Figure G16. Inter-annual variability (2002–2006) in GSI estimates of fish from the Fraser River in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.

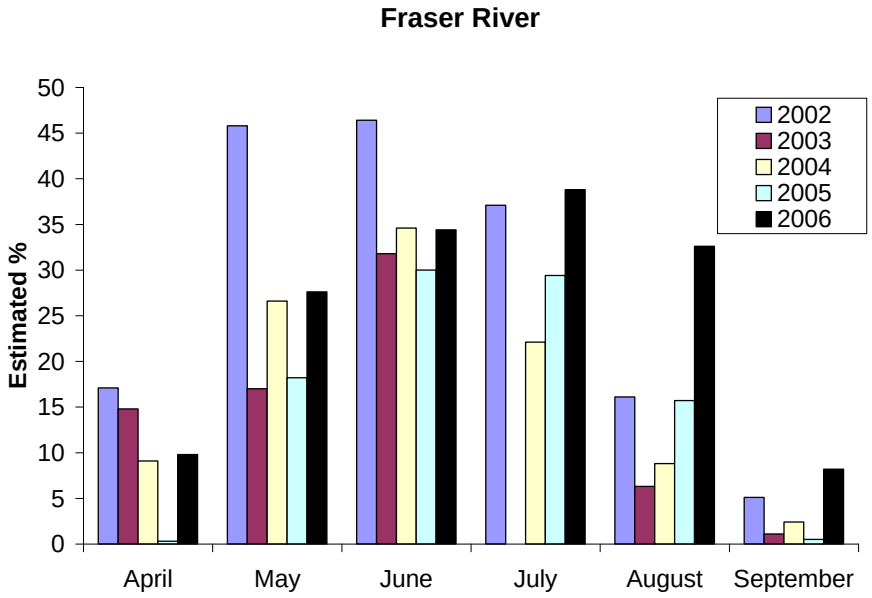

Figure G17. Inter-annual variability (2002–2006) in GSI estimates of fish from west coast of Vancouver Island drainages in Chinook salmon fishery off the northwest coast of the Queen Charlotte Islands.