kluwer
the language of science

Kluwer Academic Publishers B.V.
Van Godewijckstraat 30
3311 GX Dordrecht
The Netherlands
P.O. Box 990
3300 AZ Dordrecht
The Netherlands

Manufacturing Department
T +31 (0) 78 65 76 000
F +31 (0) 78 65 76 254
I www.wkap.nl
E services@wkap.nl

Dear Author(s),

Please find attached the PDF proof of your paper which is to be published in one of our journals. Additional forms are attached for ordering extra offprints and PDF reprints (except for Book Reviews), correcting your proof, and copyright. You will not receive a paper proof nor the original manuscript.

Note: No article can be published without a signed Consent to Publish & Transfer of Copyright form in the possession of the publisher. Please sign and return this form, along with the Offprint Order form, to the address given below.

Proof Corrections

Keep your corrections to a minimum. Only essential alterations may be made at this stage. Please use the proofs solely for checking the accuracy of the typesetting as well as the completeness and correctness of the text, tables and figures. Changes to the accepted content will not be considered.

You may find it difficult to accurately judge the figure quality due to the low resolution of the PDF. If we require better quality figures you will be contacted. If your article contains colour illustrations and you would like to receive hard copy proofs of these illustrations, please contact us at the e-mail address provided below.

Finally, proofreading is your responsibility, and although the Editor may also make corrections, Kluwer Academic Publishers will not be able to proofread the paper after it has been returned, and will not assume responsibility for any errors that you have failed to correct in the proofs.

Your response, with or without corrections, should be sent within 72 hours of receipt of this material. We ask that you do not make any corrections to the electronic file (PDF); corrections and any other comments should be submitted to KAP in one of the following ways:

Corrections must be clearly marked on a printout of the PDF file and should be sent by first class mail (airmail overseas) to:

Kluwer Academic Publishers Manufacturing Department P.O. Box 990 3300 AZ Dordrecht The Netherlands	or by courier:	Kluwer Academic Publishers Manufacturing Department Van Godewijckstraat 30 3311 GX Dordrecht The Netherlands

Only minor corrections (±10) may be sent as an e-mail attachment to: proofscorrection@wkap.nl
Always quote the four-letter journal code and number and the PIPS No. (from your proof) in the subject field of your e-mail!

The uncorrected proof of your article will be available under "Forthcoming Papers" on the journal webpage at http://journals.kluweronline.com until publication.

Your offprints are not usually available until three weeks after publication of the journal issue, when they are sent out by surface mail. Consequently, it may take 2 to 3 months for offprints to reach authors outside Europe.

We hope you are sufficiently informed.

With kind regards,

Kluwer Academic Publishers, Manufacturing Department

ELECTRONIC REPRINT ORDER FORM

After publication of your journal article, electronic (PDF) reprints may be purchased by arrangement with Kluwer Academic Publishers and Aries Systems Corporation.

The PDF file you will receive will be protected with a copyright system called DocuRights®. Purchasing 50 reprints will enable you to redistribute the PDF file to up to 50 computers. You may distribute your allotted number of PDFs as you wish; for example, you may send it out via e-mail or post it to your website. You will be able to print five (5) copies of your article from each one of the PDF reprints.

Please type or print carefully. Fill out each item completely.

1. Your name: _____

   Your e-mail address: _____

   Your phone number: _____

   Your fax number: _____

2. Journal title (vol, iss, pp): _____

3. Article title: _____

4. Article author(s): _____

5. How many PDF reprints do you want? _____

6. Please refer to the pricing chart below to calculate the cost of your order.

Number of PDF reprints	Cost (in U.S. dollars)
50	$200
100	$275
150	$325
200	$350

NOTE: Prices shown apply only to orders submitted by individual article authors or editors. Commercial orders must be directed to the Publisher.

All orders must be prepaid. Payments must be made in one of the following forms:

- a check drawn on a U.S. bank
- an international money order
- Visa, MasterCard, or American Express (no other credit cards can be accepted)

Send this form with payment to:

Aries Systems Corporation
200 Sutton Street
North Andover, Massachusetts 01845
Attn.: Electronic Reprints

PAYMENT (type or print carefully):

Amount of check enclosed: _____ (payable to Aries Systems Corporation)

❒ VISA _____

❒ MasterCard _____

❒ American Express _____

Expiration date: _____ Signature: _____

Your PDF reprint file will be sent to the above e-mail address. If you have any questions about your order, or if you need technical support, please contact: support@docurights.com

For subscriptions and to see all of our other products and services, visit the Kluwer website at :
http://www.kluweronline.com

TERMS OF DELIVERY

1. A minimum of 50 offprints may be ordered. Prices corresponding to the number of pages and quantities ordered are given below.

2. Author for correspondence will receive this offprint order form. This author is therefore also responsible for any orders the co-authors wish to make. All orders for a particular paper should appear together on only one form.

3. Offprints are printed at the same time the book or journal is printed. Thus, no alterations from the exact form in which the article appears in a book or journal are possible. Any orders for offprints that are received after the book or journal is printed should be submitted to the Production Secretariat at the address below.

4. Offprints will be forwarded within a short time after the appearance of the published paper.

5. No additional offprints can be printed or sent unless this order form is filled in, signed, returned with appropriate payment or official purchase order and received before the book or journal goes to press.

6. Any correspondence in connection with offprints must state the name of the periodical, title and offprint order number and name(s) of the author(s). In case of camera-ready publications please inform us as soon as possible, preferably by fax, about the number of offprints you wish to order.
All communications should be sent to Kluwer Academic Publishers, Manufacturing Department, P.O. Box 990, 3300 AZ Dordrecht / Van Godewijckstraat 30, 3311 GX Dordrecht, The Netherlands.
ABN-AMRO Bank, Dordrecht 50.80.13.917, Postal Cheque Account Number 4447384.

Prices of additional offprints are in EUR

Number of copies	Number of pages					
	1-4	5-8	9-12	13-16	17-20	+4
50	108	177	248	318	388	71
100	153	238	323	407	492	85
150	200	300	400	500	599	100
200	246	361	476	591	706	115
250	292	422	551	681	810	130
300	339	483	628	772	917	145
350	385	544	703	863	1022	159
400	431	606	780	955	1129	175
450	478	667	857	1046	1235	189
500	524	728	932	1136	1340	204
+50	+50	+62	+72	+92	+103	+15

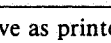
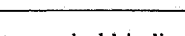
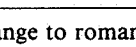
Payment will be accepted in any convertible currency. Please check the rate of exchange with your bank.
If ordering from within The Netherlands please add 6% VAT to the price quoted above.

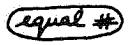
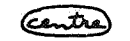
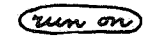
As of January 1st 1993 customers within the EEC must consider the following rules:

- If you are in possession of a VAT identification number, please fill the VAT number in on the order form. You will not be charged VAT.
- If you do *not* have a VAT number, then please add the low VAT rate - applicable to your country - to the prices quoted above.

kluwer
the language of science

PROOF CORRECTION MARKS TO BE USED BY AUTHORS

MARGINAL MARK	MEANING	CORRESPONDING MARK IN TEXT
	Delete (take out)	/ or  Cross through
	Delete and close-up	 Above and below matter to be taken out
	Leave as printed (when matter has been crossed out by mistake)	----- Under matter to remain
	Change to capital letters	Under letters or words altered
	Change to lower case letters	Encircle letters altered
	Change to bold type	 Under matter altered
	Change to bold italic type	 Under matter altered
	Change to italics	Under matter altered
	Change to roman type	Encircle matter altered
×	Replace by similar but undamaged character or remove extraneous marks	Encircle letter to be altered
	Insert (or substitute) superior figure or sign	/ or /
	Insert (or substitute) inferior figure or sign	/ or /
	Insert (or substitute) hyphen	/ or /
	Insert (or substitute) dash	/ or /
	Insert (or substitute) solidus	/ or /
...	Insert (or substitute) ellipsis	/ or /
	Close-up–delete space	 Linking words or letters

#	Insert space	| or	Between items
	Make spacing equal	|	Between items
	Reduce space	| or 	Between items
	Insert space between lines or paragraphs		
	Reduce space between lines or paragraphs		
	Transpose		Between letters or words, numbered when necessary
	Transpose lines		
	Place in centre of line		Around matter to be centred
	move to the left		
	move to the right		
	Begin a new paragraph	□	Before first word of new paragraph
	No fresh paragraph here		Between paragraphs
	(Caret mark.) Insert matter indicated in margin		
	Insert single/double quotes		

Remarks

To indicate a substitution, simply cross out the letters or words to be replaced, and write the correct letters or words in the margin. It is not necessary, nor even desirable, to use the marks for *delete* and *insert* when making a substitution. If there is more than one substitution in a line, place them in the correct order in the margin, and indicate the end of each correction with an oblique stroke / . Alternatively, continental location marks may be used, but these are to be placed in front of the corrections, not behind as in the case of the oblique stroke.

The typesetter treats *all* letters and words in the margin as insertions or substitutions, so – to avoid misunderstanding – any comments *not* intended to form part of the text should be encircled.

All alterations should be marked clearly so that there is no risk of misunderstanding; long additions or amendments should be typed on separate slips and attached. *Only really essential alterations should be made at proof stage.*

In addition to reading the proofs, please look through your edited manuscript to see if there are *any queries from the copy editor*, and if so, answer the queries *on the proofs.*

kluwer
the language of science

OFFPRINT ORDER

553534 / 5152090

Dr J.H. Reynolds
U.S. Fish and Wildlife Service
Div. of Natural Resources
1011 E. Tudor Road
Anchorage
AK 99503 Anchorage
U.S.A.

RE: **Detecting specific populations ...**
by: **REYNOLDS/TEMPLIN**

To be published in:
Environmental Biology of Fishes

Dear **Dr Reynolds**

This is to let you know that the above publication has gone into production and will appear in due course. Offprints of your article may be ordered by filling in and returning this form.

I would like to receive

50 offprints free of charge

. . . . . additional offprints without cover (minimum of 50 offprints)

Orders for offprints are only accepted if received with payment or if accompanied by an official purchase order from your institution, failing of which no offprints can be produced. Postage and handling cost are absorbed by the publishers. Payment can be made by credit card, bankdraft personal cheque or international money order. Payment is accepted in any hard currency. Prices of additional offprints and delivery terms are mentioned on the enclosed price list. Make cheques payable to KLUWER ACADEMIC PUBLISHERS – DORDRECHT

O I enclose payment to the amount of . . . . . .

O Please charge my credit card account
Card no.: . . . . . Expiry date: . . . . . .
O Access O Eurocard O America Express O Bank Americard
O Visa O Diners club O Master Charge
O I enclose official purchase order no.

O VAT identification number

Date . . . . . Signature . . . . . .

PLEASE CHECK YOUR ADDRESS AND CORRECT IF NECESSARY

kluwer
the language of science

CONSENT TO PUBLISH & TRANSFER OF COPYRIGHT

For the mutual benefit and protection of Authors and Publishers it is necessary that Authors provide formal written CONSENT TO PUBLISH and TRANSFER OF COPYRIGHT before publication of the Work.

The signed CONSENT ensures that the Publisher has the Author's permission to publish the relevant Work.

The signed TRANSFER entitles the Publisher on behalf of the Author to protect the Work against unauthorised use and to authorise dissemination of the Work by means of offprints, legitimate photocopies, microform editions, reprints, translations, and secondary information sources such as abstracting and indexing services including data bases

The Publisher hereby requests the Author to complete and return this form promptly to Kluwer Academic Publishers, Manufacturing Department, P.O. Box 990, 3300 AZ Dordrecht, The Netherlands.

Title of Contribution:
Detecting specific populations ...

Author(s):
REYNOLDS/TEMPLIN

Name of Journal:
Environmental Biology of Fishes

1. The Author hereby assigns to the Publisher the copyright to the Contribution named above whereby the Publisher shall have the exclusive right to publish the said Contribution and translations of it wholly or in part throughout the World during the full term of copyright including renewals and extensions thereafter. These rights include without limitation mechanical, electronic and visual reproduction; electronic storage and retrieval; and all other forms of electronic publication or any other types of publication including all subsidiary rights.
2. The Author retains the right to republish the Contribution in any printed collection consisting solely of the Author's own Works without charge and subject only to notifying the Publisher of the intent to do so and to ensuring that the publication by the Publisher is properly credited and that the relevant copyright notice is repeated verbatim.
3. In the event of receiving any other request to reprint or translate all or part of the Contribution the Publisher shall endeavour to obtain the approval of the Author prior to giving any such permission.
4. The Author guarantees that the Contribution is original, has not been published previously, is not under consideration for publication elsewhere, and that any necessary permission to quote from another source has been obtained. (A copy of such permission should be sent with this form.)
5. The Author declares that any person named as co-author of the Contribution is aware of the fact and has agreed to being so named.

Date . . . . . Name . . . . . .

Signature*

* To be signed by the Author, also on behalf of any co-authors, or by the Employer, where appropriate.

For Publisher's use only
Journal Issue . . . . . Pages . . . . . Ms. Ref. No. **EBFIGS2581** PIPS No. **5152090**
Year . . . . . .

Environmental Biology of Fishes **00:** 1–11, 2003.

Detecting specific populations in mixtures

Joel Howard Reynolds[a] & William David Templin
Gene Conservation Laboratory, Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, AL 99518-1599, U.S.A.
[a]*Current address: Division of Natural Resources, U.S. Fish and Wildlife Service, 1011 E. Tudor Road, Anchorage, AL 99503, U.S.A. (e-mail: Joel_Reynolds@fws.gov)*

Received 17 April 2003 Accepted 16 June 2003

Key words: component contribution, genetic stock identification, mixed stock analysis, mixture models, SPAM, species conservation

Synopsis

Mixed stock analysis (MSA) estimates the relative contributions of distinct populations in a mixture of organisms. Increasingly, MSA is used to judge the presence or absence of specific populations in specific mixture samples. This is commonly done by inspecting the bootstrap confidence interval of the contribution of interest. This method has a number of statistical deficiencies, including almost zero power to detect small contributions even if the population has perfect identifiability. We introduce a more powerful method based on the likelihood ratio test and compare both methods in a simulation demonstration using a 17 population baseline of sockeye salmon, *Oncorhynchus nerka*, from the Kenai River, Alaska, watershed. Power to detect a nonzero contribution will vary with the population(s) identifiability relative to the rest of the baseline, the contribution size, mixture sample size, and analysis method. The demonstration shows that the likelihood ratio method is always more powerful than the bootstrap method, the two methods only being equal when both display 100% power. Power declines for both methods as contribution declines, but it declines faster and goes to zero for the bootstrap method. Power declines quickly for both methods as population identifiability declines, though the likelihood ratio test is able to capitalize on the presence of 'perfect identification' characteristics, such as private alleles. Given the baseline-specific nature of detection power, researchers are encouraged to conduct *a priori* power analyses similar to the current demonstration when planning their applications.

Introduction

Mixed stock analysis (MSA) is used to estimate the relative contributions of distinct populations in a mixture of organisms. This is an important tool in fisheries management and research, with genotypes commonly used as natural markers to distinguish major populations or stocks (e.g. genetic stock identification) (Begg et al. 1999, Shaklee et al. 1999, Pearce et al. 2000). Other characteristics commonly used in fisheries include parasite assemblages (Urawa et al. 1998, Moles & Jensen 2000), scale patterns (Marshall et al. 1987), morphometrics and meristics (Fournier et al. 1984), artificial tags such as thermal marks, coded wire tags, or fin clips (Ihssen et al. 1981).

Increasingly, MSA is used to judge the presence or absence of specific populations in specific mixture samples. For example, management of an interception fishery may be heavily influenced by the presence or absence, in the harvest or bycatch, of a specific population that is threatened, weakened, or otherwise of special interest.

MSA can overestimate the contributions of populations that actually contribute little, or nothing, to a mixture (Pella & Milner 1987). Managers and researchers therefore face two questions when using MSA to judge the absence of a specific population in a mixture sample. Q1: What method should be used to test if a specific population is absent from the mixture and just receiving a nonzero estimate due to bias?

Q2: What is the method's power to detect a contribution of x% of the population from a mixture sample of size N?

Testing Population As absence is equivalent to testing for a nonzero contribution: H_0: $\theta = 0$ *versus* H_A: $\theta > 0$, where θ is Population As contribution to the mixture. The most common test assesses the one-sided lower 95% bootstrap confidence interval for θ: if the interval's limit is >0, the test rejects H_0 at a significance level of 0.05 and the population is deemed 'present'. If the interval's limit is 0, the observations present insufficient evidence to reject H_0 at 0.05 and the population contribution is deemed 'statistically indistinguishable from zero' (Seeb & Crane 1999).

AQ: Pl check ref. is not listed.

This method has a number of statistical flaws, some subtle (Reynolds & Templin in press), some more obvious. Most glaring is the method's low statistical power for detecting small contributions in application. Consider an ideal marker and an ideal population: a gene for which Population A is fixed for an allele that is unique among the other populations in the baseline, that is, a private allele. For illustration, we briefly ignore the impact of sampling the mixture and speak directly of the contribution in the mixture sample. Population A is perfectly identifiable, so a mixture sample of size N containing N times θ individuals from Population A will produce a nonzero contribution estimate, $\hat{\theta} > 0$. Even with reasonably sized mixture samples, if θ is small, say $\theta < 0.05$, then there is a positive probability that a bootstrap resample will not have any individuals from Population A; for that resample $\hat{\theta}^{\text{resample}} = \theta^{\text{resample}} = 0.0$ (Appendix 1, Part 1). The probability of this occurring increases as θ decreases to 0. Considering the roughly 1 000 resamples required for an adequate bootstrap confidence interval (Davison & Hinkley 1997, p. 156), this small probability of a resample 'without Population A' can lead to a moderate probability that the lower confidence interval will have a limit of 0 and therefore fail to detect the nonzero contribution (Appendix 1, Part 2). While this probability of a 0 lower limit is ameliorated somewhat by considering the full process – random sampling from the mixture followed by bootstrap resampling from the random sample (Appendix 1, Part 3), yet even with perfect identifiability the bootstrap confidence interval method has only little to moderate power to detect small contributions at common sample sizes (Table 1). Most importantly, as demonstrated below, the method's power is drastically reduced in application, where perfect identifiability is the rare exception. A more powerful alternative, using a likelihood ratio test, is introduced below.

We briefly review the standard MSA model and estimation method, conditional maximum likelihood estimation (Millar 1987, Pella & Milner 1987), develop the likelihood ratio test, and describe how to estimate P values using Monte Carlo simulation. The method is demonstrated in a simulation study of mixtures of sockeye salmon (*Oncorhynchus nerka*) from the Kenai River, Alaska. The likelihood ratio test and the confidence interval approach are compared in terms of their power to detect a nonzero contribution from a specific population or group of populations. The comparison demonstrates how to conduct *a priori* power analyses for a given baseline, a specific stock of interest, and a range of stock contributions and mixture sample sizes. Alternative approaches using individual assignment methods are discussed.

The likelihood ratio test is more powerful than the bootstrap confidence interval, though both methods display lower power than desired. Both methods lose power as population identifiability and contribution decline, but the bootstrap method loses power faster

Table 1. Power to detect nonzero, perfectly identifiable, contributions in MSA using the bootstrap confidence interval method, as a function of true population contribution to the original mixture (θ) and mixture sample size. In applications with less than perfect identifiability, power will be much lower (see Figure 3).

θ	Mixture sample size						
	50	75	100	150	200	300	350
0.10	0.85	0.97	1.00	1.00	1.00	1.00	1.00
0.05	0.40	0.66	0.83	0.97	0.99	1.00	1.00
0.04	0.28	0.50	0.69	0.90	0.97	1.00	1.00
0.03	0.16	0.32	0.49	0.76	0.90	0.99	1.00
0.02	0.06	0.15	0.25	0.48	0.68	0.90	0.95
0.01	0.01	0.03	0.06	0.14	0.24	0.47	0.58

Calculations are based on 1000 nonparametric bootstrap resamples and the one-sided lower 95% percentile bootstrap confidence interval (Davison & Hinkley 1997). Calculation details are in Appendix 1.

and has zero power in some scenarios. The likelihood ratio method retains a nonzero power in all scenarios. In the absence of perfect identifiability, the likelihood ratio method's low power limits application to those problems involving small to moderate-sized baselines of potentially contributing populations.

Methods

The finite mixture model

The following model describes mixtures of contributions from finitely many source populations (see, e.g. Millar 1987 or Pella & Milner 1987). Although the presentation assumes discrete characteristics are observed on each individual, such as a genotype, this is not essential; the model holds for continuous characteristics as well.

AQ: Pls check symbol '≥'.

Let N individuals be randomly sampled from a mixture of J populations. Let the jth population contribute an unknown proportion $\theta_j \geq 0$ to the mixture, $\Sigma\theta_j = 1$; $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_J)$. If the characteristic measured on the ith sample individual is denoted by x_i, then the probability of observing the sample $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is

$$\begin{aligned} \Pr(\mathbf{X} \mid \boldsymbol{\Theta}, \boldsymbol{\Pi}) &= \prod_{i=1}^{n} \Pr(x_i \mid \boldsymbol{\Theta}, \boldsymbol{\Pi}) \\ &= \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} \theta_j \Pr(x_i \mid \boldsymbol{\pi}_j) \right\} \end{aligned} \tag{1}$$

where $\boldsymbol{\pi}_j$ is a column vector of parameters specifying the probability density function for the characteristic in population j, and $\boldsymbol{\Pi}$ is the matrix $[\boldsymbol{\pi}_1 | \cdots | \boldsymbol{\pi}_J]$. For a discrete characteristic with k possible outcomes, $\boldsymbol{\pi}_j = (\pi_1^j, \ldots, \pi_k^j)$, $\pi_i^j \geq 0, \sum_i \pi_i^j = 1$, the vector of multinomial probabilities. This assumes that the set {Pop.1, . . ., Pop. J} includes all potentially contributing populations (see Smouse et al. 1990). Expanding the $\Pr(x_i \mid \boldsymbol{\pi}_j)$ terms allows for multivariate characteristics.

Estimation

Estimating the mixture proportions, $\boldsymbol{\Theta}$, requires information regarding the (possibly multivariate) characteristic probability density function, $\boldsymbol{\pi}_j$, for each contributing population. This is generally available in the form of a sample from each baseline population. In most fisheries applications researchers fix the nuisance parameters, $\boldsymbol{\pi}_j$, at their estimates from the baseline samples, $\hat{\boldsymbol{\pi}}_j$ (Millar 1987). Maximum likelihood is then used to estimate the unknown $\boldsymbol{\Theta}$ conditional on $\boldsymbol{\pi}_j = \hat{\boldsymbol{\pi}}_j$. This conditioning is justified by the small amount of information on $\boldsymbol{\pi}_j$ in the mixture sample, relative to the baseline sample (Milner et al. 1981). Recently developed Bayesian methods utilize this information, which may be an important consideration when analyzing related mixture samples collected through space or time (Pella & Masuda 2001).

Identifiability of a stock in a mixture requires, among other things (Pella & Milner 1987), that the probability density functions of the characteristics, $\boldsymbol{\pi}_j$, differ across the contributing populations (Redner & Walker 1984). Characteristics commonly used in fisheries include parasite assemblages (Urawa et al. 1998, Moles & Jensen 2000), scale patterns (Marshall et al. 1987), morphometrics and meristics (Fournier et al. 1984), artificial tags such as thermal marks, coded wire tags, or fin clips (Ihssen et al. 1981), and increasingly, genetic markers (Seeb & Crane 1999, Ruzzante et al. 2000).

Uncertainty in the mixture proportion estimates, $\hat{\boldsymbol{\Theta}}$, arises from sampling uncertainty in both the mixture and the population baselines samples. In practice, these sampling uncertainties can be accounted for by nonparametric bootstrap resampling from the mixture sample and parametric bootstrap resampling from the baseline characteristic distributions, $\hat{\boldsymbol{\pi}}_j$. Bootstrap resampling of the baseline samples increases the width of the resulting confidence intervals, reducing the power to detect nonzero contributions. The following demonstration only resamples the mixture sample.

Testing population absence

Assume a sample is taken from a mixture consisting of contributions from a known set of baseline populations, with specific interest in testing the absence of Population A, H_0: $\theta_A = 0$. The likelihood ratio test compares the likelihood of the observed sample under the general model, in which Population A contributes, to the likelihood under the null model, in which Population A does not contribute (that is, $\theta_A = 0$). The likelihood ratio test statistic, conditional on $\boldsymbol{\pi}_j = \hat{\boldsymbol{\pi}}_j$, is

$$\begin{aligned} \mathrm{LR} &= \frac{L(\{\theta_1, \theta_2, \ldots, \theta_J\} \mid \mathbf{X}, \hat{\boldsymbol{\Pi}})}{L(\{\theta_1', \theta_2', \ldots, \theta_A' = 0, \ldots, \theta_J'\} \mid \mathbf{X}, \hat{\boldsymbol{\Pi}})} \\ &= \frac{\prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} \theta_j \Pr(x_i \mid \hat{\boldsymbol{\pi}}_j) \right\}}{\prod_{i=1}^{n} \left\{ \sum_{\substack{j=1 \\ j \neq A}}^{J} \theta_j' \Pr(x_i \mid \hat{\boldsymbol{\pi}}_j) \right\}} \end{aligned} \tag{2}$$

with $\{\theta_1, \ldots, \theta_J\}$ and $\{\theta'_1, \ldots, \theta'_J\}$ replaced by their conditional maximum likelihood contribution estimates under their respective models. The observed ratio, LR^{obs}, is calculated by fitting the mixture sample using the full baseline (the general model, the numerator), then fitting the mixture sample using the reduced baseline with Population A dropped to force $\theta_A = 0$, (the null model, the denominator). The test can be extended to the joint contribution of a specific group of populations, H_0: $\theta_{A1} = \theta_{A2} = \cdots \theta_{Av} = 0$ *versus* H_a: one or more of $\{\theta_{A1}, \ldots, \theta_{Av}\} > 0$.

If Population A has a unique characteristic, relative to the other baseline populations, that also occurs in the mixture sample, the likelihood under the reduced baseline will be zero, giving a likelihood ratio of ∞. Commonly used MSA software (Debevec et al. 2000) assigns such individuals to an 'unknown' baseline component, clearly identifying the nonzero contribution of Population A.

AQ: Pls check if the symbol '∞' is correct.

Completing the test requires comparing the observed ratio, LR^{obs}, to its expected distribution under H_0. There are two ways to estimate this distribution.

Null reference distribution method 1: asymptotic theory

In theory, the null hypothesis can be tested by comparing -2 times $\ln(LR^{obs})$ to its asymptotic distribution under the null model, a χ^2 with degree of freedom equal to the number of populations being simultaneously tested for zero contribution (Stuart et al. 1999). The conditions underlying this asymptotic result do not hold when other populations in the baseline fail to contribute to the mixture (Stuart et al. 1999). As this is often the case in fisheries genetic stock identification problems (Millar 1987), the asymptotic results are frequently unreliable. Even when all populations are expected to have nonzero estimates, experience has shown that the asymptotic results may remain unreliable.

Null reference distribution method 2: Monte Carlo simulation

The null reference distribution can be approximated by Monte Carlo simulation under H_0, conditional on $\boldsymbol{\Theta}_0$ (Davison & Hinkley 1997, p. 138). The value of $\boldsymbol{\Theta}_0$ will generally not be known prior to analysis and must be estimated from fitting the null model. Estimate $\boldsymbol{\Theta}_0$ from the observed mixture sample by fitting model (1) using the reduced baseline, giving $\hat{\boldsymbol{\Theta}}_0$. Then simulate R sets of N observations from model (1) using the estimated null mixture proportions, $\hat{\boldsymbol{\Theta}}_0$, and the baseline population characteristic densities $\hat{\boldsymbol{\Pi}}$. Take each set of simulated observations and fit model (1) using the full baseline (Population A included), giving an estimate $\hat{\boldsymbol{\Theta}}^{*r}$. Take each set of simulated observations and fit model (1) using the reduced baseline (no Population A included), giving an estimate $\hat{\boldsymbol{\Theta}}_0^{*r}$. For each r, calculate and record the likelihood ratio (Equation (2)), LR^{*r}.

This process gives a sample of size R, $\{LR^{*r}: r = 1, \ldots, R\}$, from the unknown null reference distribution. Calculate the observed likelihood ratio, LR^{obs}, by fitting the general and restricted models to the actual mixture sample and plugging the estimates into Equation (2). An approximate P value for the test is given by $\left(1 + \sum_r I(LR^{*r} \geq LR^{obs})\right)/(1 + R)$ (Davison & Hinkley 1997, p. 141), where the indicator function I( ) takes the value one when the argument is true and zero otherwise. Generally, R in the range 1000–5000 provides sufficient precision (Davison & Hinkley 1997, sec. 4.2.5).

Uncertainty in the conditional values of the nuisance parameters, $\hat{\boldsymbol{\pi}}_j$, can be incorporated into the Monte Carlo simulation approach by parametric bootstrap resampling from each $\hat{\boldsymbol{\pi}}_j$ before constructing the null mixture during each of the R simulation rounds. The resulting null reference distribution will actually be a mixture of null reference distributions, one for each resampled baseline. This resampling will increase the dispersion in the null reference distribution and hence potentially overestimate the tail areas of interest. For ease of comparison, the following demonstration does not include resampling of the baseline estimates.

Demonstration

The two methods were compared in terms of their power to detect a nonzero population contribution. The simulation study used an allozyme baseline of 19 markers for the sockeye salmon populations of Kenai River, Alaska (collection and analysis details in Seeb et al. 2000, nomenclature following Shaklee et al. 1990): *mAAT-1**, *mAAT-2**, *mAH-1,2**, *mAH-4**, *sAH**, *ALAT**, *G3PDH-1,2**, *GPI-A**, *GPI-B1,2**, *sIDHP-1**, *LDH-B2**, *sMDH-A1,2**, *PEPA**, *PEPC**, *PEPB-1**, *PEPD-1**, *PEPLT**, *PGM-1**, *PGM-2**. To investigate how population identifiability influences detection power, we explored three population sets of declining, though relatively high, identifiability.

Kenai River sockeye baseline

The Kenai River is the major producer of sockeye salmon in Cook Inlet, Alaska, supporting a commercial fishery in the inlet, a personal use fishery at the river mouth, and a recreational fishery within the river itself (Figure 1). The river fisheries are managed to allow a set range of individuals to reach the spawning grounds. Resource managers are interested in detecting the presence of specific populations at different times during the fishing season.

The Kenai River baseline consists of seventeen populations. Geography (Figure 1) and genetic diversity (Figure 2), followed by extensive simulation analyses, were used to define five regions, or population aggregates, that were reliably identified in mixture estimation (Seeb et al. 2000). A region had to demonstrate a contribution estimate of 90% or more, averaged over 500 bootstrap resamples, from simulated mixture samples of 400 genotypes, where each mixture sample was drawn uniformly from the populations within the region of interest, and each genotype was parametrically bootstrapped from its population of origin's allele frequency estimates. Five regions were identified (Figures 1 and 2): Upper Russian River (two populations), Hidden Creek, Trail Lakes (three populations), Tern Lake, and Kenai/Skilak (10 populations). Three scenarios were investigated: detecting the highly identifiable Upper Russian River region, the moderately identifiable Trail Lakes region, and the somewhat less identifiable Tern Lake region.

The Upper Russian River drainage occurs above a waterfall, which acts as a partial barrier to upstream movement. The populations spawning above the falls are relatively genetically distinct (Seeb et al. 2000), though they do not exhibit any private alleles at the allozyme markers considered. The Railroad Creek population in the Trail Lakes region exhibits a private allele at *mAH-1,2** (relative frequency 0.013); the Tern Lake population exhibits a private allele at *mAH-4** (relative frequency 0.01).

Simulated mixtures

For each reporting region of interest, mixture samples of genotypes from 200 individuals were simulated over a range of population contributions, $\theta_{\text{Region\_of\_interest}} = \{10\%, 5\%, 4\%, 3\%, 2\%, 1\%\}$. For Upper Russian River or Trail Lakes, $\theta_{\text{Region\_of\_interest}}$ was evenly split among the region's populations. The remaining populations in the baseline evenly contributed the rest of the mixture. A contribution from a given population was simulated by randomly generating a genotype from that population's allele frequencies

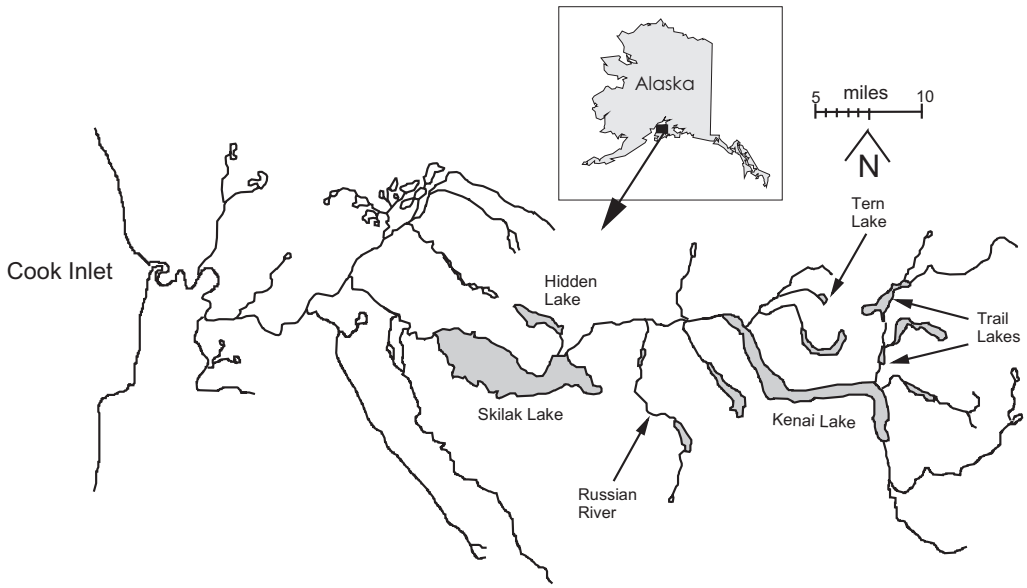


Figure 1. Major sockeye salmon producing lake systems of the Kenai River watershed, Cook Inlet, Alaska. Each labeled lake system constitutes a reporting region in the baseline (Figure 2) with the exception of Skilak and Kenai Lakes, which are combined into one region.

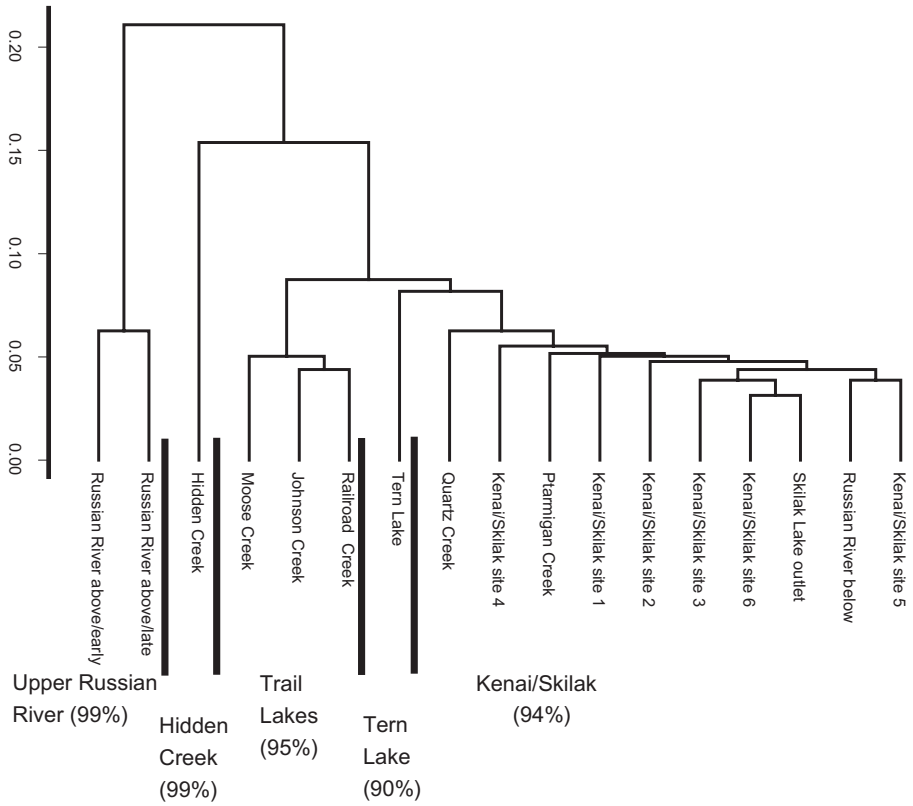


Figure 2. UPGMA tree of populations in the Kenai River sockeye salmon baseline, using Cavalli-Sforza and Edwards genetic distance (described in Weir 1996) on 19 allozyme markers (see text for details). The populations are aggregated into five reporting regions for MSA, where a reporting region is the smallest set of populations that achieves, on average, a 90% or greater contribution estimate for simulated mixtures generated strictly of individuals from the populations themselves ('100% simulations') (see Seeb et al. 2000 for details). The region labels also give the mean contribution estimate for each region's 100% simulations. The Kenai/Skilak sites are located along the reach between the two lakes (Figure 1).

for each of the 19 allozyme markers. Baseline allele frequencies are available in Seeb et al. (2000). Fifty mixture samples were simulated for each combination of region of interest and contribution level.

Analyses

Each mixture sample was analyzed to estimate: the reporting region contributions under both the full and reduced baseline models, the one-sided 95% lower percentile bootstrap confidence interval for $\theta_{\text{Region\_of\_interest}}$ under the full baseline model, and all quantities required to conduct the likelihood ratio test of H_0: $\theta_{\text{Region\_of\_interest}} = 0$ *versus* H_A: $\theta_{\text{Region\_of\_interest}} > 0$. The bootstrap confidence interval used $B = 1000$ resamples; for equal numerical accuracy in tail estimation, the Monte Carlo approximation to the null reference distribution used $R = 1000$ simulations.

The bootstrap and likelihood ratio test were compared in terms of their power to detect the nonzero contribution of the region of interest. For a given scenario – (method and region of interest and contribution), power was estimated as the percentage of the 50 simulated mixture samples for which the method detected a nonzero contribution from the region of interest. Detection was defined as: bootstrap method – nonzero limit lower limit on the one-sided 95% confidence interval when rounded to two significant digits; likelihood ratio – P value ≤ 0.05 or nonzero contribution assignment to the 'unknown' category when fitting the mixture using the reduced baseline model. Baseline allele frequencies were not resampled.

Mixture samples were generated using S-Plus 2000 (Insightful, Inc., Seattle, WA, U.S.A.) and locally written programs. Mixture analyses were conducted using the freeware package SPAM 3.5 (Reynolds 2001).

Demonstration results

The likelihood ratio test was as or more powerful than the bootstrap confidence interval method in testing population absence (Figure 3), detecting at least every contribution the bootstrap method detected. Equality occurred only when both methods displayed 100% power. Both methods displayed less than 'ideal' power (Figure 3 vs. Table 1 for the bootstrap confidence interval method; Figure 3 vs. 100% for the likelihood ratio method). The likelihood ratio method always displayed positive power.

Even under the most optimistic scenario (perfect identifiability), the discrete nature of bootstrap resampling drives the confidence interval method's detection power to near zero for very small contributions, $\theta \leq 2\%$, and common sample sizes (Table 1). In practice, the method displays a drastically reduced power that quickly drops to very low levels for even relatively sizeable contributions, $\theta \leq 5\%$, for any populations with less than extremely high identifiability (Figure 3).

Both methods lost power as the region of interest's identifiability declined or the true contribution declined (Figure 3). The likelihood ratio method's power did not decline as quickly as that of the bootstrap confidence interval method (Figure 3).

Discussion

Increased usage of MSA has increased demand for methods of detecting small nonzero contributions from a specific population, and hence for distinguishing, to the extent possible, such contributions from mere biased estimates for an absent population. The likelihood ratio test is a more powerful method than the current bootstrap confidence interval approach. It detected every contribution the bootstrap method detected, and

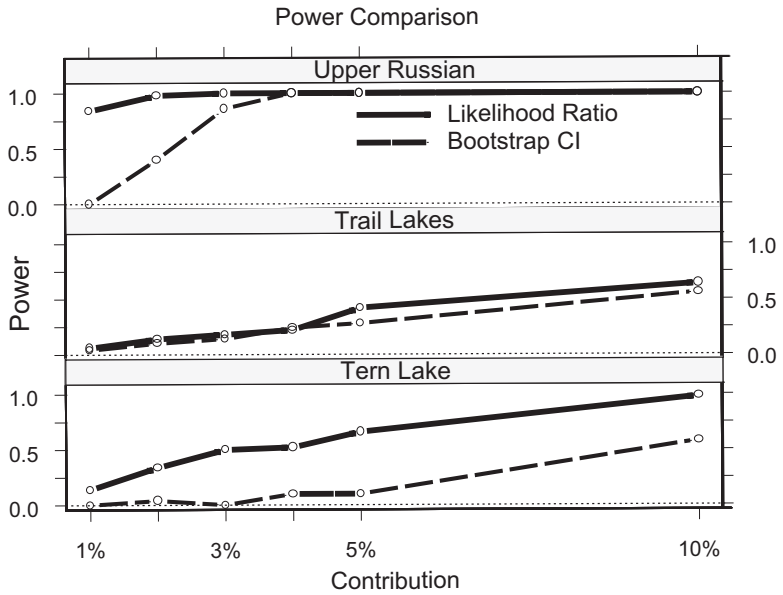


Figure 3. Power to detect nonzero contributions, by method, for each investigated region of interest. The likelihood ratio test displays higher power than the bootstrap confidence interval method for detecting a specific region's contribution, regardless of which specific region (panel), or contribution level (%) (x axis). The likelihood ratio is able to maintain ideal power to detect even two individuals out of 200 (Russian River panel). While both methods lose power with decreasing identifiability of the population(s) of interest (refer to Figure 2), the likelihood ratio method is able to retain a positive, albeit small, power at every scenario; the power of the bootstrap method quickly goes to zero in all but the high identifiability situations. Fifty mixtures of 200 individuals were simulated for each contribution level by region of interest combination; see text for details. Power estimates are the proportion of each scenario's 50 simulated mixtures in which the method detected the contribution.

more (Figure 3), while maintaining a positive detection power in every scenario.

The increase in power of the likelihood ratio test over the bootstrap confidence interval is not surprising. The bootstrap's usefulness for generating confidence intervals stems from its almost complete absence of parametric assumptions. In the case of MSA, however, the parametric assumptions associated with the mixture model and Hardy–Weinburg equilibrium are generally not contentious. Accepting these assumptions, which is already implicit in using standard methods of MSA estimation, lets one employ traditional parametric methods of parameter testing and their associated gain in power.

Further, using the bootstrap confidence interval for a parameter test actually entails more subtle statistical assumptions that, in this application, are known to be false. One must assume the contribution estimate is a pivotal statistic (see Lunneborg 2000), and one must accept the arithmetic difference as a useful measure of distance between compositional data vectors (see Aitchison 1992).

The likelihood ratio test's increase in power, relative to the bootstrap confidence interval method, varies with both the identifiability of the population(s) of interest relative to the rest of the baseline and the number of observations the population actually contributes to the mixture sample. When the populations in a region of interest are highly identifiable, both methods are capable of 100% power even for moderately small contributions (Russian River, Figure 3). In these situations, the likelihood ratio method retains 100% power at every contribution level.

As the population(s) of interest becomes increasingly similar to other members of the baseline, the few individuals actually contributed by the population(s) of interest may be adequately explained as having originated from the similar populations. When these similar populations are themselves contributing to the mixture, such as during the Tern Lake scenarios where the Kenai/Skilak region contributes a relatively large portion of the mixture, the principle of parsimony underlying the likelihood ratio test will lead to absorption of the individuals from the population(s) of interest into that component contributed by the similar populations. For example, at low-contribution levels, the few observations from Tern Lake appear to be easily absorbed into the already substantial Kenai/Skilak region contribution.

This absorption is avoided if the population(s) of interest has either sufficiently distinct characteristics that allow clear detection of even a single contribution (e.g. private alleles or the Russian River scenario – Figures 2 and 3), or has a large enough contribution to the mixture sample such that the likelihood ratio test is able to detect the signal in the sample's joint distribution of characteristics (Figure 3). It is sobering to see how easily a contribution can be absorbed, that is, how low the power to detect a nonzero contribution can be (Figure 3, Trail Lakes). The power demonstrated here, with a baseline of only 17 populations, may decrease even more as baseline size increases.

Normal standards of mixture analysis identifiability, that is, the performance on 100% simulations, do not appear to directly relate to expected power in detecting small contributions. For example, relative to Trail Lakes, Tern Lake has a lower correct contribution estimate from its 100% simulations yet it appears to exhibit a higher power of detectability when testing for population absence. This may be an artifact of the simulation method: the Trail Lakes contribution is simulated as equal portions from all three component populations. Relative to the single population contribution from Tern Lake, this partitioning of the Trail Lakes contribution lessens any population-specific contribution, weakening the regional signal.

The likelihood ratio test is very sensitive to private alleles due to the conditional likelihood estimation method. If one is confident that the private allele is, indeed, a population-specific marker, then this makes for a very sensitive test. However, if the private allele is likely an artifact of limited sampling of highly polymorphic markers, this sensitivity may mislead. A substantial portion of the Trail Lakes and Tern Lake contribution detections involved such rare alleles (Table 2). The current demonstration does not allow us to judge whether or not a contribution would have

Table 2. Detections of contribution as a result of private alleles (left) compared with the total detections of contribution using the likelihood ratio method (right), out of 50 simulations per scenario.

Region	Contribution (%)					
	10	5	4	3	2	1
Trail lakes	13/32	11/21	4/11	5/9	0/7	1/3
Tern lake	22/49	6/33	8/26	9/25	6/17	0/7

The likelihood ratio method detects the presence of even a single individual, regardless of mixture sample size, if that individual displays a characteristic unique to its source population. The likelihood ratio method detected every contribution from the Upper Russian River, even though it has no private alleles.

been detected in the absence of a rare allele, that is, strictly by the signal inherent in the joint distribution of characteristics contributed by the population(s) of interest. The Russian River results clearly demonstrate that the likelihood ratio method does not require private alleles to detect contributions (Figure 3). The impact on detection of a given 'private allele' would be revealed by repeating the analysis with the allele in question recoded as a common allele, that is, by binning alleles.

One may consider abandoning MSA for this question and using an analysis that directly assigns each observation in the mixture sample to the 'most probable' population of origin. This can lead to a less powerful test. Conditional maximum likelihood-based individual identification methods (e.g. Cornuet et al. 1999, Banks & Eichert 2000) analyze each observation independently. Since they ignore the information in the mixture sample's joint distribution of characteristics, the methods cannot provide a more powerful test of absence than the MSA likelihood ratio method, and generally will be less powerful. In contrast, Bayesian methods of MSA that estimate, for each observation and each population, the probability the observation came from that population, do utilize the information in the mixture sample's joint distribution of characteristics (Pella & Masuda 2001). For the question of interest, such Bayesian methods will likely have comparable power to the MSA likelihood ratio.

The Bayesian method provides two other appealing features for the question under consideration. First, Bayesian MSA will provide a direct estimate of the distribution of any specific population's mixture contribution, θ. This allows the researcher to directly assess the Prob($\theta > 0$) for that population. Further, one can explore whether a large Prob($\theta > 0$) arises because of a couple of highly distinguishable observations, or simple due to the collective weight of a large number of observations each of which displays some possibility of having originated from the population of interest. Explorations of this application of Bayesian methods are currently in progress.

AQ: Pls check symbol 'θ'.

Regardless of which method one utilizes to assess population absence, one should repeat the demonstration process illustrated here to assess the method's power. Such *a priori* analyses allow one to determine the sample size required to detect a given contribution with a given power, as well as compare methods in a specific context. Note that posterior power analyses, while unfortunately rather common, are generally uninformative and, therefore, misleading (Hoenig & Heisey 2001). *A priori* power analysis methods in the general context of MSA are discussed elsewhere (Reynolds 2001).

Acknowledgements

We are indebted to P.A. Crane for motivating discussions, L.W. Seeb for support and encouragement, and Sandra Sonnichsen for Figure 1. The manuscript, and the authors' thinking on this topic, were greatly enhanced by the contributions of two anonymous reviewers. Contribution PP-227 of the Alaska Department of Fish and Game, Commercial Fisheries Division, Juneau.

References

Aitchison, J.A. 1992. On criteria for measures of compositional difference. Math. Geol. 24: 365–379.

Banks, M.A. & W. Eichert. 2000. WHICHRUN (Version 3.2) a computer program for population assignment of individuals based on multilocus genotype data. J. Hered. 91: 87–89.

Begg, G.A., K.D. Friedland & J.B. Pearce. 1999. Stock identification – its role in stock assessment and fisheries management. Fish. Res. 43: 1–3.

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

Davison, A.C. & D.V. Hinkley. 1997. Bootstrap Methods and their Application, Cambridge University Press, Cambridge, UK. 582 pp.

Debevec, E.M., R.B. Gates, M. Masuda, J.J. Pella, J.H. Reynolds & L.W. Seeb. 2000. SPAM (version 3.2): Statistics Program for Analyzing Mixtures. J. Hered. 91: 509–511.

Fournier, D.A., T.D. Beacham, B.E. Ridell & C.A. Busack. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. Can. J. Fish. Aquat. Sci. 41: 400–408.

Hoenig, J.M. & D.M. Heisey. 2001. The abuse of power: The pervasive fallacy of power calculations of data analysis. Amer. Stat. 55: 19–25.

Ihssen, P.E., H.E. Booke, J.M. Casselman, J.M. McGlade, N.R. Payne & F.M. Utter. 1981. Stock identification: Materials and methods. Can. J. Fish. Aquat. Sci. 38: 1838–1855.

Lunneborg, C.E. 2000. Data Analysis by Resampling: Concepts and Applications, Duxbury Press, Pacific Grove, CA, U.S.A. 568 pp.

Marshall, S., D. Bernard, R. Conrad, B. Cross, D. McBride, A. McGregor, S. McPherson, G. Oliver, S. Sharr & B. Van Allen. 1987. Application of scale patterns analysis to

the management of Alaska's sockeye salmon (*Onchorhynchus nerka*) fisheries. pp. 207–326. *In*: H.D. Smith, L. Margolis & C.C. Wood (ed.) Sockeye Salmon (*Oncorhynchus nerka*) Population Biology and Future Management, Canadian Special Pub. on Fish. and Aquat. Sci. 96.

Millar, R.B 1987. Maximum likelihood estimation of mixed stock fishery composition. Can. J. Fish. Aquat. Sci. 44: 583–590.

Milner, G.B., D.J. Teel, F.M. Utter & C.L. Burley. 1981. Columbia River stock identification study: Validation of genetic method. Final report of research (FY80) financed by Bonneville Power Administration Contract DE-A179-80BP18488, National Marine Fisheries Service, Northwest and Alaska Fisheries Center, Seattle, WA. 35 pp.

Moles, A. & K. Jensen. 2000. Prevalence of the sockeye salmon brain parasite *Myxobolus arcticus* in selected Alaska streams. Alaska Fish. Res. Bull. 6: 85–93.

Pearce, J.M., B.J. Pierson, S.L. Talbot, D.V. Derksen, D. Kraege & K.T. Scribner. 2000. A genetic evaluation of morphology used to identify harvested Canada geese. J. Wild. Manag. 64: 863–874.

Pella, J.J. & M. Masuda. 2001. Bayesian methods for stock-mixture analysis from genetic characters. Fish. Bull. 99: 151–167.

Pella, J.J. & G.B. Milner. 1987. Use of genetic marks in stock composition analysis. pp. 247–276. *In*: N. Ryman & F. Utter (ed.) Population Genetics and Fishery Management, Washington Sea Grant Program, Seattle, WA.

Redner, R.A. & H.F. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. Soc. Indust. Appl. Math. Rev. 26: 195–239.

Reynolds, J.H. 2001. SPAM version 3.5: User's guide addendum. Addendum to special publication No. 15, Alaska Dept. of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab, 333 Raspberry Rd., Anchorage, AK, 99518–1599 (available: www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm).

Ruzzante, D.E., C.T. Taggart, S. Lang & D. Cook. 2000. Mixed-stock analysis of Atlantic cod near the Gulf of St. Lawrence based on microsatellite DNA. Ecol. Appl. 10: 1090–1109.

Seeb, L.W. & P.A. Crane. 1999. Allozymes and mitochondrial NDA discriminate Asian and North American populations of chum salmon in mixed-stock fisheries along the south coast of the Alaska Peninsula. Trans. Amer. Fish. Soc. 128: 88–103.

Seeb, L.W., C. Habicht, W.D. Templin, K.E. Tarbox, R.Z. Davis, L.K. Brannian & J.E. Seeb 2000. Genetic diversity of sockeye salmon of Cook Inlet, Alaska, and its application to management of populations affected by the *Exxon Valdez* Oil Spill. Trans. Amer. Fish. Soc. 129: 1223–1249.

Shaklee, J.B., F.W. Allendorf, D.C. Morizot & G.S. Whitt. 1990. Gene nomenclature for protein-coding loci in fish. Trans. Amer. Fish. Soc. 119: 2–15.

Shaklee, J.B., T.D. Beacham, L. Seeb & B.A. White. 1999. Managing fisheries using genetic data: Case studies from four species of Pacific salmon. Fish. Res. 43: 45–78.

Smouse, P.E., R.S. Waples & J.A. Tworek. 1990. A genetic mixture analysis for use with incomplete source population data. Can. J. Fish. Aquat. Sci. 47: 620–634.

Stuart, A., J.K. Ord & S. Arnold. 1999. Kendall's Advanced Theory of Statistics Vol 2A: Classical Inference and the Linear Model, 6th edition, Oxford University Press, New York, NY, U.S.A. 885 pp.

Urawa, S., K. Nagasawa, L. Margolis & A. Moles. 1998. Stock identification of chinook salmon (*Onchorhynchus tshawytscha*) in the North Pacific Ocean and Bering Sea by parasite tags. Nor. Pacific Anadro. Fish. Com. Bull. 1: 199–204.

Weir, B.S. 1996. Genetic data analysis II, Sinauer Associates, Sunderland, MA, U.S.A. 445 pp.

Appendix 1

Ideal power to detect nonzero contributions using bootstrap confidence intervals.

Part 1. Assume that the population of interest, Population A, contributes a proportion $0 < \theta < 1$ of uniquely identifiable individuals to the mixture and that N individuals are randomly and independently sampled from the mixture. For example, consider red balls (Population A) randomly mixed with blue balls (Population B), in proportion θ to $1 - \theta$, in an infinite barrel. N balls are randomly selected as the mixture sample. A one-sided lower 95% confidence interval is constructed from 1000 bootstrap resamples using the percentile method (Lunneborg 2000).

Let k be the number of individuals from Population A in the original mixture sample, $k \sim$ binomial(N, $p = \theta$). Then Prob(no Population A individuals in one resample $|\, k$) = Prob(no red balls in random sample, with replacement, of original N balls) $= (1 - k/N)^N = \nu$.

AQ: Pls check symbol 'ν'.

Part 2. Let X be the number of times out of 1000 resamples in which no red balls are found. X is a binomial random variable with $N = 1000$ trials and 'success' probability $= \nu$. Then the probability a confidence interval calculated using the one-sided 95% percentile method includes 0 is the probability that 50 or more of the resamples will contain no red balls, i.e. that $X \geq 50$.

AQ: Pls check symbol '$\geq$'.

$$\text{Prob}(X \geq 50) = 1 - \text{Prob}(X < 50)$$
$$= 1 - \sum_{X=0}^{49} \binom{1000}{X} \nu^X (1-\nu)^{1000-X}$$

So Prob(CI calculated using one-sided 95% percentile method does not include 0)

$$\text{Prob}(\text{X} < 50)$$

$$= 1 - \left[1 - \sum_{\text{X}=0}^{49} \binom{1000}{\text{X}} \nu^{\text{X}} (1-\nu)^{1000-\text{X}}\right]$$

$$= \sum_{\text{X}=0}^{49} \binom{1000}{\text{X}} \nu^{\text{X}} (1-\nu)^{1000-\text{X}}$$

Part 3. The power to detect Population As nonzero contribution θ to the mixture, using the bootstrap confidence interval method, is

$$\sum_{\text{k}=0}^{\text{N}} (\text{Power} \mid \text{k}, \text{N}) \text{Prob}(\text{K} = \text{k} \mid \theta, \text{N})$$

$$= \sum_{\text{k}=0}^{\text{N}} \left(\sum_{\text{X}=0}^{49} \binom{1000}{\text{X}} \nu^{\text{X}} (1-\nu)^{1000-\text{X}} \right)$$

$$\times \text{Prob}(\text{K} = \text{k} \mid \theta, \text{N})$$

$$= \sum_{\text{k}=0}^{\text{N}} \left(\sum_{\text{X}=0}^{49} \binom{1000}{\text{X}} \left(\left(1 - \frac{\text{k}}{\text{N}}\right)^{\text{N}} \right)^{\text{X}} \left(1 - \frac{\text{k}}{\text{N}}\right)^{\text{N}} \right)^{1000-\text{X}}$$

$$\times \binom{\text{N}}{\text{k}} \theta^{\text{k}} (1-\theta)^{\text{N}-\text{k}}$$