Dividing population genetic distance data with the software Partitioning Optimization with Restricted Growth Strings (PORGS): an application for Chinook salmon (*Oncorhynchus tshawytscha*), Vancouver Island, British Columbia

John R. Candy (contact author)[1]

R. Gregory Bonnell[2]

Terry D. Beacham[1]

Colin G. Wallace[1]

Ruth. E. Withler[1]

Email address for contact author: John.Candy@dfo-mpo.gc.ca

[1] Molecular Genetics Laboratory
Department of Fisheries and Oceans, Pacific Biological Station
3190 Hammond Bay Road
Nanaimo, British Columbia, Canada V9T 6N7

[2] Oceans and Habitat Enhancement Branch
Department of Fisheries and Oceans
4166 Departure Bay Road
Nanaimo, British Columbia, Canada V9T 4B7

Abstract—A new method of finding the optimal group membership and number of groupings to partition population genetic distance data is presented. The software program Partitioning Optimization with Restricted Growth Strings (PORGS), visits all possible set partitions and deems acceptable partitions to be those that reduce mean intracluster distance. The optimal number of groups is determined with the gap statistic which compares PORGS results with a reference distribution. The PORGS method was validated by a simulated data set with a known distribution. For efficiency, where values of n were larger, restricted growth strings (RGS) were used to bipartition populations during a nested search (bi-PORGS). Bi-PORGS was applied to a set of genetic data from 18 Chinook salmon (*Oncorhynchus tshawytscha*) populations from the west coast of Vancouver Island. The optimal grouping of these populations corresponded to four geographic locations: 1) Quatsino Sound, 2) Nootka Sound, 3) Clayoquot +Barkley sounds, and 4) southwest Vancouver Island. However, assignment of populations to groups did not strictly reflect the geographical divisions; fish of Barkley Sound origin that had strayed into the Gold River and close genetic similarity between transferred and donor populations meant groupings crossed geographic boundaries. Overall, stock structure determined by this partitioning method was similar to that determined by the unweighted pair-group method with arithmetic averages (UPGMA), an agglomerative clustering algorithm.

Manuscript submitted 26 March 2008.
Manuscript accepted 5 September 2008.
Fish. Bull. 107:45–56 (2009).

Genetic diversity in salmon species is thought to be maintained through high homing fidelity, which limits gene flow between spawning sites (Ricker, 1972; Quinn and Dittman, 1990). As a general rule, populations that are geographically close tend to be genetically similar, creating natural clusters of similar populations. Identification of genetically similar salmonid populations is important for fisheries management initiatives directed at conserving genetic diversity (Riddell, 1993; Waples et al., 2001). Consequently, managers are faced with the challenge of defining the number and size of these genetic groups. Furthermore, determining valid groupings of populations at a fine scale allows managers to make informed decisions regarding harvest levels and population-enhancement strategies. For British Columbia Chinook salmon (*Oncorhynchus tshawytscha*) populations, genetic markers have been used to determine genetic distance between populations and to provide considerable power for defining regional stock structure (Teel et al., 2000; Beacham et al., 2006a).

Clustering or grouping data are useful in many disciplines; as a result there is a wide assortment of methods available for representing data, measuring proximity between data elements, and grouping elements (e.g., Jain et al., 1999). For Pacific salmon, population-specific allelic frequencies are ascertained from spawning ground samples by using genetic markers at a number of loci. From these allelic frequencies, a metric of overall genetic difference between populations is used to estimate pairwise genetic distances. Three commonly used distance measures are Nei's distance, D_S (Nei, 1987), Nei's modified Cavalli-Sforza chord distance D_A (Cavalli-Sforza and Edwards, 1967; Nei et al., 1983), and Weir and Cockerham's (1984) estimator of F_{st}, the coancestory coefficient θ. Once a distance measure is selected, a proximity matrix is created which shows genetic distance between each pair of populations.

Clustering is often used to group populations, either by merging small clusters into larger ones (agglomerative) or by splitting larger clusters

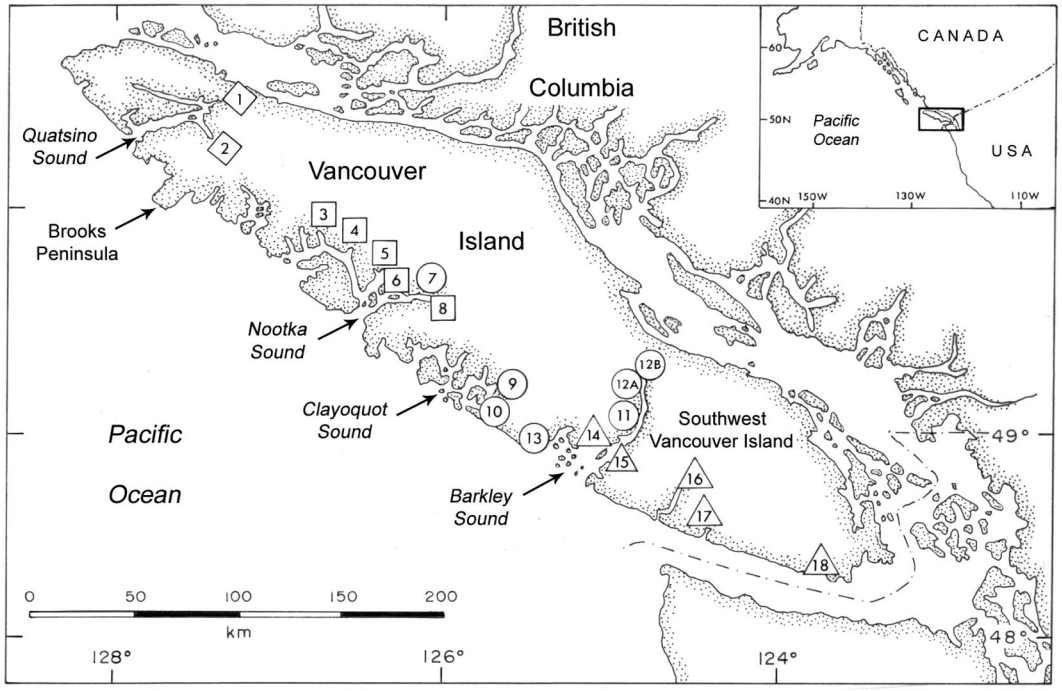


Figure 1

Location of the 18 sites on the west coast of Vancouver Island where Chinook salmon (*Oncorhynchus tshawytscha*) populations were sampled. Numbers correspond to stock codes in Table 1. The same population was sampled for Somas River (12A) and Robertson Hatchery (12B). Shapes around location numbers denote an genetic affiliation with one of the four regional groups: Quatsino Sound (diamonds), Nootka Sound (squares), Clayoquot+Barkley aounds (circles), and southwest Vancouver Island (triangles).

into smaller ones (divisive). A number of algorithms are available to decide which small clusters are merged or which larger clusters are split (e.g., Swofford et al., 1996; Jain et al., 1999). Groupings can be depicted as a branching tree or dendrogram where branch length is scaled to represent genetic distance. A drawback with the hierarchical approach is that the result is sensitive to initial groupings, which are not permitted to change once an assignment has been made. Furthermore, arbitrary tie-breaking actions, either in the original proximity data or during agglomeration, can cause instability in the tree structure (van der Kloot et al., 2005). Consensus from multiple tree constructions by bootstrapping across loci provides a measure of robustness of the apparent dominant tree structure (Felsenstein, 1985). A majority-rule consensus tree can provide a phylogeny with groups that occur in a majority of the bootstrap samples. However, the incorporation of variation from consensus trees appears to have limited quantitative application, and the optimum cluster number is not obvious.

This article provides a new method for partitioning genetic distance data by finding the optimal group membership and number of groupings. We validate the method using simulated data. To demonstrate the utility of this partition method, we applied it to genetic distance data calculated from samples taken from 18 Chinook salmon populations along the west coast of Vancouver Island, British Columbia (Fig. 1). The groupings determined by this method were evaluated with respect to known transfers of broodstock and histories of stock enhancement. Furthermore, results from both the simulated and Chinook salmon data sets were compared to results from a commonly used clustering method for genetic data.

Materials and methods

Pairwise cost function

A pairwise cost function used in the field of pattern recognition (Roth et al. 2003) minimizes the sum of mean intracluster distances. Minimized intracluster distance appears most desirable in grouping populations where two or more populations assigned to the same group contribute to total cost. Other clustering algorithms have been proposed which emphasize separation, combinations of compactness and separation, or conductivity measures (Buhmann, 2002).

Given row (i) and column (j) indices of an ($n \times n$) dissimilarity matrix D of populations with k groups, the pairwise cost function (CF) is

$$CF = \frac{1}{2}\sum_{v=1}^{k}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}M_{iv}M_{jv}D_{ij}}{\sum_{l=1}^{n}M_{lv}}. \quad (1)$$

For each population a binary assignment variable indicates group membership such that group membership (l) is assigned to each group (v) in an ($n \times k$) binary matrix (M), where

$$M \in \{0,1\}^{nxk} : \sum_{v=1}^{k} M_{iv} = 1. \quad (2)$$

The optimal assignments of $\hat{M}$ are obtained through cost-function minimization ($\downarrow CF$) and visiting all combinations of group memberships. Unlike other cost functions (Hofmann and Buhmann, 1997), there is no penalty for increased numbers of partitions; thus adding more partitions will always reduce the cost where the output is nonconvex and $CF \to 0$ as $k \to n$. A nonpenalizing cost function was implemented so that the gap statistic (discussed later) could be used for determining the optimal number of groupings. Adding partitions creates more and smaller groups while lowering mean intracluster distance (the sum of all pairwise distances divided by the number of populations). Meanwhile, adding more groups increases the sum of the mean intracluster distances.

Implementation of the search algorithm

Testing all group memberships at different cluster sizes can generate large numbers of combinations. Structure detection through partitioning is considered a combinatorial optimization problem because visits to all combinations are computationally intensive. There is no guarantee of finding the optimal solution in a reasonable amount of time because the number of computations grows rapidly with increasing data (Puzicha et al., 1999). We describe two search methods that have been used for these data: simple random search and complete search.

Simple random search, a random set-partition assignment of the binary matrix, is an obvious way to visit combinations of group memberships, where i = 1 to n such that

$$M(i, rand\_v) = 1. \quad (3)$$

Alternatively, all $n \times k$ combinations can be visited as a complete list of set partitions where, for example, three populations can be partitioned into the form

ABC AB|C AC|B A|BC A|B|C.

Set partitions are the union of nonempty disjoint subsets called blocks, where restricted growth strings (RGS) (strings of numbers used as a convenient way to represent partitions) were used to generate all blocks (Knuth, 2005). We called visits to all partitions while minimizing the cost function (Eq. 1), partitioning optimization using restricted growth strings (PORGS). The number of ways n populations can be partitioned into these nonempty sets is called the Bell number (Rota, 1964; Cameron, 1994). The total number of set partitions is the n^{th} Bell number, and the number of set partitions for each k is determined by the Stirling number of the second kind (Cameron, 1994).

Set partitions determined by RGS were used to configure the binary matrix to assign group membership. Although RGS can visit all possible partitions, they can also be used to generate partitions with "at most" r blocks (Knuth, 2005). This reduced search space allows bipartition (bi-PORGS) (r=2) such that an optimum split can be determined one partition at a time. Information from prior group membership is used to restrict future searches, where

$$M(i, v) = 1 \text{ for } i = 1 \text{ to } l, \text{ where } v = 1 \text{ or } 2. \quad (4)$$

A nested search occurs when all subgroups are sorted in descending order, and block combinations are selected when the cost function is minimized. Computational search time is reduced with the bi-PORGS method, thus allowing partitioning of larger sets of data.

The gap statistic

The objective of this analysis was to find an optimum number of groups, as well as the optimum partition solution, for k groups. Although there is no one criterion for deciding how many groups should be chosen to best represent the data, one guiding principle is that the appropriate number occurs when additional groups do not substantially change within-cluster dispersion. The gap statistic reveals within-cluster dispersion with that expected under an appropriate reference null distribution with methods of Tibshirani et al., (2001) such that

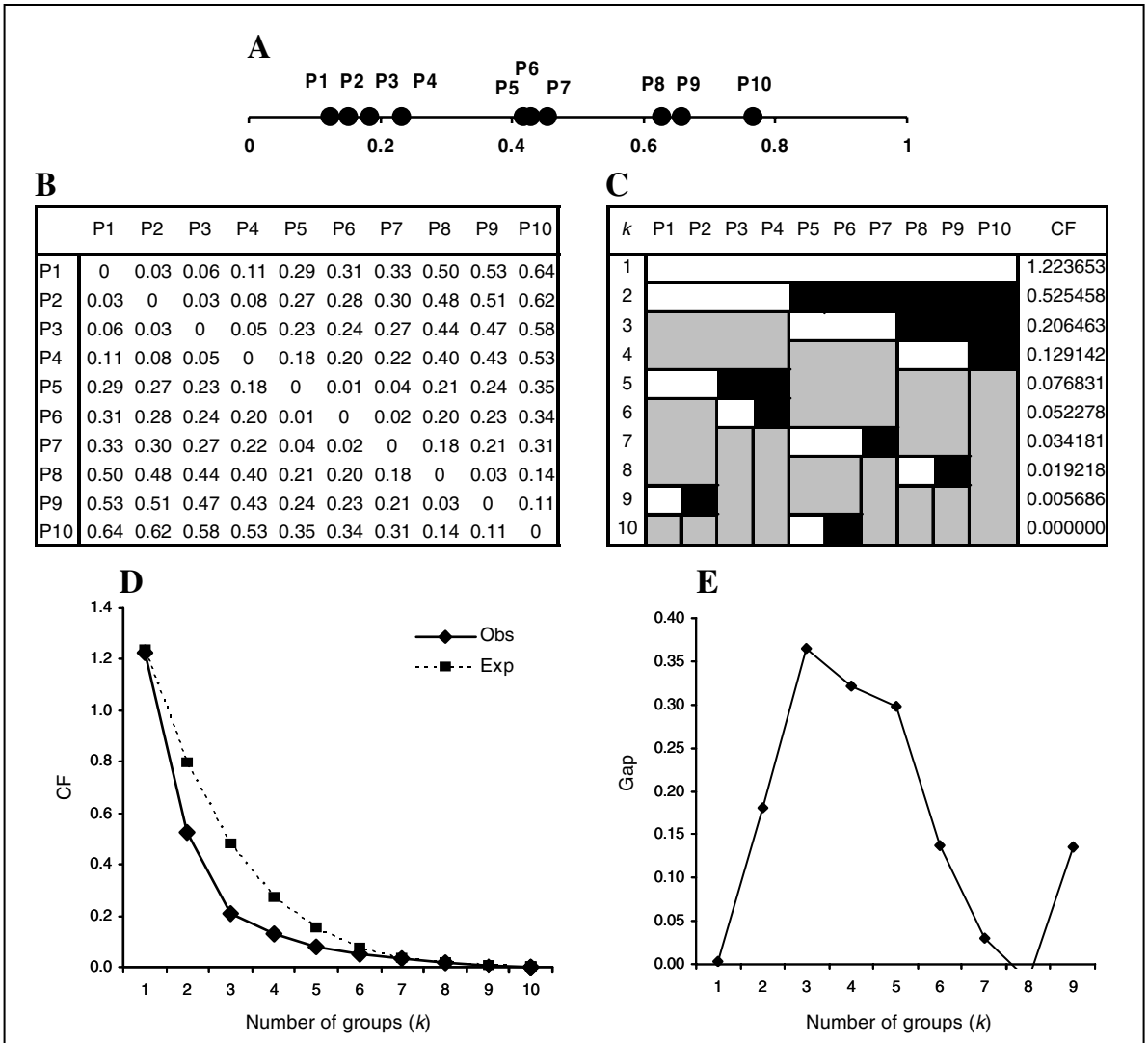
$$Gap_n(k) = \mathrm{E}^*_n\{\log(\downarrow CF)\} - \log(\downarrow CF), \quad (5)$$

where $\downarrow CF$ = the observed values from the minimized cost function for each k; and
$\mathrm{E}^*_n\{\log(\downarrow CF)\}$ = the log of the expected values from the reference distribution for each k.

The gap statistic is largest when the observed values fall the farthest below the reference curve. The estimate of the optimum number of groups will be the value where additional groups do not increase the gap statistic. The expected values for the reference distribution are generated by taking the mean PORGS values from bootstrapping the proximity matrix. Essentially, the mean values from the bootstrapped matrices remove the stock structure component from the reference data.

Simulated data

Simulated data were used to validate the PORGS method by comparing the known distribution of data points with



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	0	0.03	0.06	0.11	0.29	0.31	0.33	0.50	0.53	0.64
P2	0.03	0	0.03	0.08	0.27	0.28	0.30	0.48	0.51	0.62
P3	0.06	0.03	0	0.05	0.23	0.24	0.27	0.44	0.47	0.58
P4	0.11	0.08	0.05	0	0.18	0.20	0.22	0.40	0.43	0.53
P5	0.29	0.27	0.23	0.18	0	0.01	0.04	0.21	0.24	0.35
P6	0.31	0.28	0.24	0.20	0.01	0	0.02	0.20	0.23	0.34
P7	0.33	0.30	0.27	0.22	0.04	0.02	0	0.18	0.21	0.31
P8	0.50	0.48	0.44	0.40	0.21	0.20	0.18	0	0.03	0.14
P9	0.53	0.51	0.47	0.43	0.24	0.23	0.21	0.03	0	0.11
P10	0.64	0.62	0.58	0.53	0.35	0.34	0.31	0.14	0.11	0

Figure 2

Partitions of simulated genetic distance data. (**A**) Populations P1–P10 are located randomly along a line from zero to one; (**B**) simulated proximity data were generated by using line distance between populations P1–P10; (**C**) results of the partitioning optimization using restricted growth strings (PORGS) analysis of simulated data, showing hierarchical group membership for k groupings and corresponding minimized cost function (CF) values; black, gray, and white represent status of the populations for a given cluster number. Black and white represent populations involved in partitions for a particular value of k, whereas gray populations were not involved; (**D**) expected CF values from the reference distribution and observed values plotted against number of groups k for the simulated data; and (**E**) the gap statistic plotted against number of groups, showing k=3 as the optimum number of groups with the simulated data for 10 populations.

optimal partitions and group membership obtained from the model.

To simulate genetic distance data, ten populations were assumed to be randomly located along a horizontal line, where i = 1 to 10 for a population $P(i)$ selected at random (Fig. 2A).

$$0 < P(i) < 1. \quad (6)$$

A proximity matrix of i rows and j columns of simulated data was constructed from the distances d between two populations (Fig. 2B) such that the values of each matrix element were

$$d(ij) = |P(i) - P(j)|. \quad (7)$$

Using the PORGS method, we partitioned the proximity matrix into optimal group membership for k = 1 to n groups by minimizing the cost function (Eq. 1; Fig. 2C). Ten bootstrapped variants of the proximity matrix were treated similarly, and the mean of the bootstrapped values provided a reference distribution to compare with

Table 1

Chinook salmon (*Oncorhynchus tshawytscha*) sampling data by region, stock code, population, years sampled, and annual and total sample size used in the bi-partitioning optimization using restricted growth strings (bi-PORGS) analysis. (H) indicates major hatchery facilities.

Region	Stock code	Population	Years sampled	Annual *n*	Total *n*
Quatsino Sound	1	Marble	1994, 1996, 1999, 2000	58, 98, 149 202	507
	2	Colonial	1999, 2004	40, 18	58
Nootka Sound	3	Zeballos	2002, 2004	4, 30	34
	4	Tahsis	1996, 1999, 2002, 2003	72, 87, 104 47	310
	5	Conuma (H)	1988, 1996, 1997, 1998	47, 214, 143 52	456
	6	Tlupana	2002, 2003	34, 32	66
	7	Gold	1983, 1985, 1986	9, 13, 71	93
	8	Burman	1976, 1985, 1986, 1989, 1990, 1991, 1992, 2000, 2002, 2003	8, 20, 2, 35 19, 56, 35 34, 51, 13	273
Clayoquot Sound	9	Tranquil	1996, 1999	209, 133	342
	10	Kennedy	1992, 2005	49, 190	239
Barkley Sound	11	Nahmint	1996, 2001, 2002, 2003, 2004	27, 56, 51 124, 135	346
	12A	Somas	1973	155	155
	12B	Robertson (H)	1996, 2003	155, 183	338
	13	Thornton	1992, 1999, 2000, 2001	37, 147, 150 184	518
	14	Toquart	1999, 2000	70, 17	87
	15	Sarita	1996, 1997, 2001	112, 157, 146	415
Southwest Vancouver Island	16	Nitinat (H)	1989, 1996, 2003	53, 153, 140	346
	17	San Juan	2001, 2002	80, 116	196
	18	Sooke	2004	58	58

the observed values of the original proximity matrix by using the gap statistic (Eq. 5; Fig. 2, D and E).

Genetic data for Chinook salmon from the west coast of Vancouver Island

Next we applied PORGS to genetic data derived from Chinook salmon populations. Tissue or scale samples were taken from Chinook salmon during times of broodstock collection and spawner enumeration along the west coast of Vancouver Island (Fig. 1). Genomic DNA was isolated from samples taken at 18 sites since the early 1970s (Table 1) and polymerase chain reaction analysis was conducted to amplify 12 microsatellite markers (*Ogo2, Ogo4, Oke4, Oki100, Ots100, Ots101, Ots104, Ots107, Ots2, Ots9, Omy325,* and *Ssa197*) by standard methods (Beacham et al., 2006a). All samples were combined over years except for one sample from the Robertson Creek Hatchery which was used to test the temporal stability of the microsatellites for population differentiation. Above the confluence with Sproat River, the Somas becomes Stamp River. Robertson Creek Hatchery lies on Stamp River downstream from Great Central Lake. The sample collected from Somas River in 1973 was considered to be the same stock as that represented by the sample from Robertson Creek Hatchery taken in 1996 and 2003. The Robertson Creek Hatchery stock was founded from Somas River fish collected from 1972 to 1976. and from additional broodstock collected from the river since then (Table 1, stock codes 12a and 12b).

For this analysis, we used Weir and Cockerham's (1984) co-ancestory coefficient θ, a widely used measure of genetic differentiation (Waples and Gaggiotti, 2006). A pairwise estimate of θ was calculated from data on multilocus genotypic distance between populations by using FSTAT software (Goudet, 1995).

History of Chinook salmon brookstock from the west coast of Vancouver Island

The west coast of Vancouver Island has three major hatcheries, Robertson Creek, Nitinat River, and Conuma River, which began enhancement of Chinook salmon in 1972, 1990, and 1979, respectively (Cross et al., 1991). Juveniles from these hatcheries have been transplanted into several river systems (Table 2). Robertson Creek Hatchery provided the founder stock for Thornton Creek Hatchery with transfers from brood years 1982–84. Since then, the hatchery-supported run has been perpetuated by returns to Thornton Creek. The Nitinat River Hatchery has transferred Nitinat River Chinook salmon to Toquart River (broods 1990–97 and 1999–2001) and

Table 2

History of transfers of Chinook salmon (*Oncorhynchus tshawytscha*) from the west coast of Vancouver Island, showing for each recipient population whether there was an indigenous stock, the enhancement status since transfers, the donor population, and the brood years transferred from donor to recipient sites.

Recipient population	Indigenous stock	Enhancement status	Donor population	Brood years transferred
Colonial	Yes	Not enhanced	Marble	1987–1989, 1991,1993,1998
Thornton	No	Enhanced: river returns only	Robertson	1982–1984
Toquart	Yes		Robertson	1989–1990
		Enhanced: river returns only	Nitinat	1990–1997, 1999–2001
Sooke	Yes	Enhanced: transfers plus river returns	Nitinat	1980–1984, 1987–1997, 1999–2006
Zeballos	Yes	Enhanced: river returns only	Conuma	1990–1998, 1999–2003

to Sooke River (brood years 1980–2006, except 1995, 1986, and 1998. Both Toquart River and Sooke River had existing Chinook salmon runs before hatchery releases, but populations were at a low abundance before transfers began. In addition, Marble River stock has been transferred to Colonial Creek, Goodspeed River (not sampled), and Coal Harbour, as well as various seapen release locations in Quatsino Sound. Other populations were either enhanced on-site or reared in a hatchery before being returned to the natal stream for release. For example, the Sarita River stock was reared in the Nitinat Hatchery, then returned to the Sarita River.

Comparison of PORGS method with standard genetic methods

To compare results obtained by the PORGS method with those from a standard hierarchical approach to clustering genetic data, the unweighted pair-group method using arithmetic averages (UPGMA) was applied to both the simulated and Chinook salmon data sets to generate a tree with PHYLIP software (Felsenstein, 1989). The UPGMA approach uses successive agglomeration with average-linking (Sneath and Sokal, 1973).

Results

When PORGS was applied to the simulated data, the first partition (k=2) occurred between P4 and P5 (Fig. 2C). The next partition occurred between P7 and P8; with k = 4, P10 was separated from the P7–P9 grouping. The last two populations to split were P5 and P6. The values of the optimized cost function decreases monotonically as the number of groups k increased (Fig. 2D). The expected data from the reference distribution (bootstrapped proximity matrix) also decreased monotonically (Fig. 2D); but followed a less concave curve than that for the observed data. The optimum k value occurred where the observed data fell the farthest below the expected curve, at k = 3 (Fig. 2E), which corresponded to the groupings P1–P4, P5–P7, and P8–P10 (Fig. 2C). According to the relative positions of these populations along the line in Figure 2A, the group memberships and optimum number of groups determined by PORGS appears reasonable. These data were re-analyzed by using the bi-PORGS method, which generated the same cluster groupings as PORGS.

These groupings are also consistent with the results depicted by an UPGMA tree (Fig. 3A). Figure 3A shows a vertical dashed line drawn to intersect branches that correspond to the three main clusters identified in the PORGS analysis. The UPGMA tree shows P10 with the longest branch length and P5–P6 with the shortest branch length, corresponding to both population location along the line (Fig. 2A) and the results from the application of the PORGS method (Fig. 2C).

The structure of the Chinook salmon data was evident when populations were sorted as an anti-Robinson matrix (Fig. 4; Robinson, 1951). In Figure 4 the smallest dissimilarity values appeared close to the main diagonal, resulting in a grouping of the most similar populations. Four main clusters were apparent, and the two Quatsino Sound populations, Marble River and Colonial Creek, were the most distinctive of all the populations (θ>0.04). Three other groups lay along the main diagonal, where θ<0.02, corresponding to northern (Nootka Sound), central (Clayoquot+Barkley sounds), and southwest Vancouver Island. The Toquart River and the Sarita River populations, although geographically part of Barkley Sound, clustered with southwest Vancouver Island populations. The Gold River population straddled the northern and central populations. The most genetically similar samples (θ=0.002) were those from Somas River in the early 1970s and those from the more recently sampled Robertson Creek hatchery.

Using the Chinook salmon data set, we were unable, in an initial attempt, to find an optimum solution from random set partitions. Although some partitions occurred only a few times, others were more common (Fig. 5). After 5.0×10^8 iterations, no cluster combinations were evaluated where block sizes were less than 2 or greater than 16. The expected number of occurrences for each k (Stirling number of the second kind)

indicated that considerably more iterations would be required to visit all set partitions.

Compared with random set partitions, PORGS reduces the number of cost function evaluations by eliminating redundant cluster combinations. However, if PORGS passes through all set partitions of the Chinook salmon data set, it would generate 5.8×10^{12} evaluations of the cost function (19th Bell number). Compared to the simulated data, where n=10, the Chinook salmon data set requires a much larger search (Fig. 5). Evidently, the number of combinations grows rapidly with an increasing number of populations, but not as fast as n factorial. (Knuth, 2005). An exhaustive search would still take a long time to execute, even with recent advances in computer processing speed. However, with bi-PORGS, the largest bipartition occurred for the first cluster with n=19 members, where 262,142 evaluations were generated.

During the bi-PORGS analysis for each value of k, the cost function minimizes the mean intracluster distance, then sums the means across all clusters. Because we were using the co-ancestory coefficient, θ, as a distance measure for the Chinook salmon data set, minimized cost function was referred to as the "mean sum theta" ($\Sigma\bar{\theta}$). For k=1 to 19, bi-PORGS values ($\downarrow CF$) represent optimal membership for each of these groups (Fig. 6A). The two most northerly Chinook salmon populations, Marble River and Colonial River (within the Quatsino Sound grouping), formed the first bipartition when ($\Sigma\bar{\theta}$=0.23). When k=4 ($\Sigma\bar{\theta}$=0.11) the remaining single cluster divided into 1) Nootka Sound, 2) Clayoquot+Barkley sounds, and 3) southwest Vancouver Island groups. Next, San Juan separated from the southwest Vancouver Island group,and the Quatsino group of Marble River and Colonial River split at k=6 ($\Sigma\bar{\theta}$=0.076). When eight groupings were optimized ($\Sigma\bar{\theta}$= 0.046) the Robertson-Creek–derived populations separated from Gold River and Nahmint River populations, as well as the Clayoquot Sound populations; and Sarita River populations split from the southwest Vancouver Island populations. The Burman River population split from the Nootka Sound populations at k=10 ($\Sigma\bar{\theta}$=0.029), and Clayoquot Sound populations (Tranquil River and Kennedy River) separated from the Barkley Sound populations (Gold River and Nahmint River). At k=12 ($\Sigma\bar{\theta}$=0.019) Thornton Creek split from Robertson Creek and Somas River, and the Gold and Nahmint Rivers split apart. At k=14 ($\Sigma\bar{\theta}$=0.011) Sooke River split from the Nitinat River and Toquart River populations, and Tranquil River and Kennedy River populations split. The last few remaining splits separated Tahsis River and Conuma River, Nitinat River and Toquart River, and finally Somas River and Robertson Creek populations (Fig. 6A).

As with the simulated data, the relationship between the number of groups and bi-PORGS evaluations decreases monotonically with increasing values of k (Fig. 6B); however, unlike the simulated situation, there appears to be more than one optimal point. The gap statistic indicates that the first optimum num-

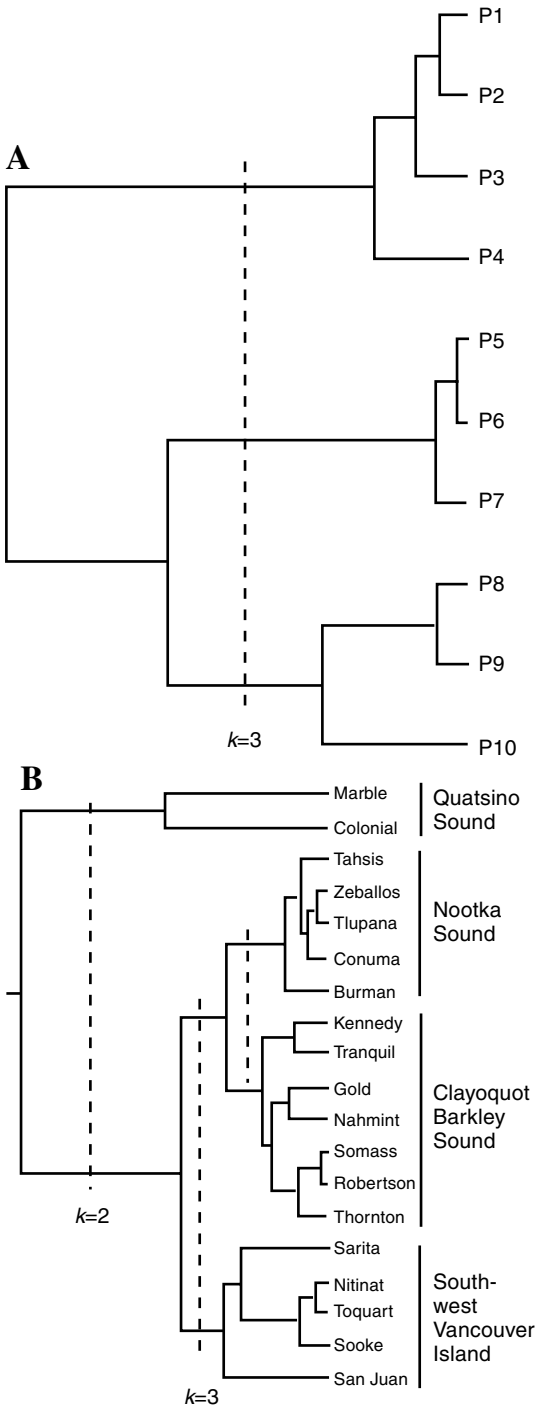


Figure 3

Dendrograms derived from (**A**) simulated genetic distance data clustered by the unweighted pair-group method using arithmetic averages (UPGMA) and from (**B**) genetic distance data for 18 populations of Chinook salmon (*Oncorhynchus tshawytscha*) from the west coast Vancouver Island clustered by using the co-ancestory coefficient θ. The dotted lines indicate corresponding groupings determined by bi-partitioning optimization using restrictive growth strings (bi-PORGS).

	Quatsino Snd		Nootka Sound					Clayoquot + Barkley sounds							Southwest Vancouver Island				
Population	Marble	Colonial	Zeballos	Tahsis	Conuma	Burman	Tlupana	Kennedy	Tranquil	<Gold>	Nahmint	Somas	Robertson	Thornton	San Juan	<Sarita>	<Toquart>	Nitinat	Sooke
Marble	0.000	0.040	0.050	0.053	0.059	0.061	0.058	0.051	0.048	0.040	0.036	0.037	0.043	0.052	0.054	0.063	0.069	0.072	0.076
Colonial	0.040	0.000	0.100	0.096	0.106	0.101	0.103	0.096	0.094	0.086	0.075	0.083	0.087	0.098	0.086	0.103	0.111	0.108	0.116
Zeballos	0.050	0.100	0.000	0.006	0.007	0.011	0.003	0.017	0.017	0.009	0.015	0.023	0.023	0.029	0.035	0.024	0.033	0.033	0.032
Tahsis	0.053	0.096	0.006	0.000	0.008	0.010	0.005	0.023	0.019	0.014	0.016	0.027	0.028	0.036	0.033	0.029	0.033	0.034	0.037
Conuma	0.059	0.106	0.007	0.008	0.000	0.013	0.003	0.027	0.024	0.018	0.025	0.034	0.035	0.043	0.046	0.034	0.042	0.041	0.044
Burman	0.061	0.101	0.011	0.010	0.013	0.000	0.008	0.022	0.021	0.014	0.019	0.031	0.031	0.040	0.042	0.029	0.038	0.038	0.043
Tlupana	0.058	0.103	0.003	0.005	0.003	0.008	0.000	0.027	0.025	0.017	0.023	0.032	0.035	0.042	0.044	0.035	0.042	0.043	0.043
Kennedy	0.051	0.096	0.017	0.023	0.027	0.022	0.027	0.000	0.009	0.014	0.015	0.018	0.017	0.021	0.031	0.027	0.037	0.037	0.042
Tranquil	0.048	0.094	0.017	0.019	0.024	0.021	0.025	0.009	0.000	0.014	0.015	0.015	0.014	0.021	0.030	0.021	0.034	0.035	0.044
Gold	0.040	0.086	0.009	0.014	0.018	0.014	0.017	0.014	0.014	0.000	0.010	0.009	0.011	0.019	0.033	0.027	0.035	0.036	0.034
Nahmint	0.036	0.075	0.015	0.016	0.025	0.019	0.023	0.015	0.015	0.010	0.000	0.013	0.013	0.019	0.027	0.024	0.030	0.033	0.037
Somas	0.037	0.083	0.023	0.027	0.034	0.031	0.032	0.018	0.015	0.009	0.013	0.000	0.002	0.008	0.030	0.032	0.044	0.046	0.046
Robertson	0.043	0.087	0.023	0.028	0.035	0.031	0.035	0.017	0.014	0.011	0.013	0.002	0.000	0.007	0.030	0.026	0.039	0.040	0.043
Thornton	0.052	0.098	0.029	0.036	0.043	0.040	0.042	0.021	0.021	0.019	0.019	0.008	0.007	0.000	0.036	0.030	0.043	0.046	0.053
San Juan	0.054	0.086	0.035	0.033	0.046	0.042	0.044	0.031	0.030	0.033	0.027	0.030	0.030	0.036	0.000	0.030	0.022	0.023	0.027
Sarita	0.063	0.103	0.024	0.029	0.034	0.029	0.035	0.027	0.021	0.027	0.024	0.032	0.026	0.030	0.030	0.000	0.018	0.016	0.029
Toquart	0.069	0.111	0.033	0.033	0.042	0.038	0.042	0.037	0.034	0.035	0.030	0.044	0.039	0.043	0.022	0.018	0.000	0.003	0.007
Nitinat	0.072	0.108	0.033	0.034	0.041	0.038	0.043	0.037	0.035	0.036	0.033	0.046	0.040	0.046	0.023	0.016	0.003	0.000	0.008
Sooke	0.076	0.116	0.032	0.037	0.044	0.043	0.043	0.042	0.044	0.034	0.037	0.046	0.043	0.053	0.027	0.029	0.007	0.008	0.000

Figure 4

Dissimilarity matrix of θ values for genetic distance data derived from Chinook salmon (*Oncorhynchus tshawytscha*) populations from the west coast of Vancouver Island. Values of θ where $\theta = 0$ are shown as white, $\theta < 0.02$ as light gray, $\theta < 0.04$ as dark gray, and $\theta \geq 0.04$ as black. Populations that show genetic affiliation but are outside the geographic region are denoted by < >.

ber of groups occurs at $k = 4$, corresponding to four geographic locations: Quatsino Sound, Nootka Sound, Clayoquot+Barkley sounds, and southwest Vancouver Island (Fig. 6C). At this point, additional groups do not cause the observed data to continue to drop substantially below the reference distribution. When k=9, a second peak in the optimum groupings occurred which corresponded to the partitioning of Barkley Sound and Clayoquot Sound populations, similar groupings were derived from the UPGMA tree; the vertical line in Figure 3B indicates the corresponding number of groups and group membership as determined by biPORGS. The four regional groups, Quatsino Sound, Nootka Sound, Clayoquot+Barkley sounds, and southwest Vancouver Island, each formed a cluster on the UPGMA tree; however, the optimum number of clusters is not obvious. The dendrogram appears to show greater genetic distance between the populations of Marble River and Colonial River than between populations of Nootka Sound and Clayoquot+Barkley sounds. Overall, the partition and agglomerative methods produced similar results.

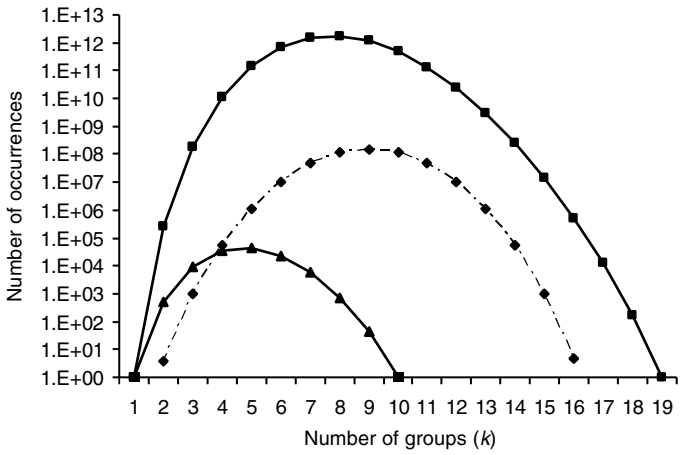


Figure 5

Number of occurrences by k groups for the simulated data set (—▲—) and for the Chinook salmon (*Oncorhynchus tshawytscha*) data set (n=19) with random search and 5.0×10^8 iterations (----◆----); and the total number of all set partition occurrences required for 19 populations (Stirling numbers of second kind (—■—).

Discussion

This article provides a new method for clustering genetic distance data by partitioning optimally with RGS, where acceptable partitions reduce intracluster distance. For this analysis, we used

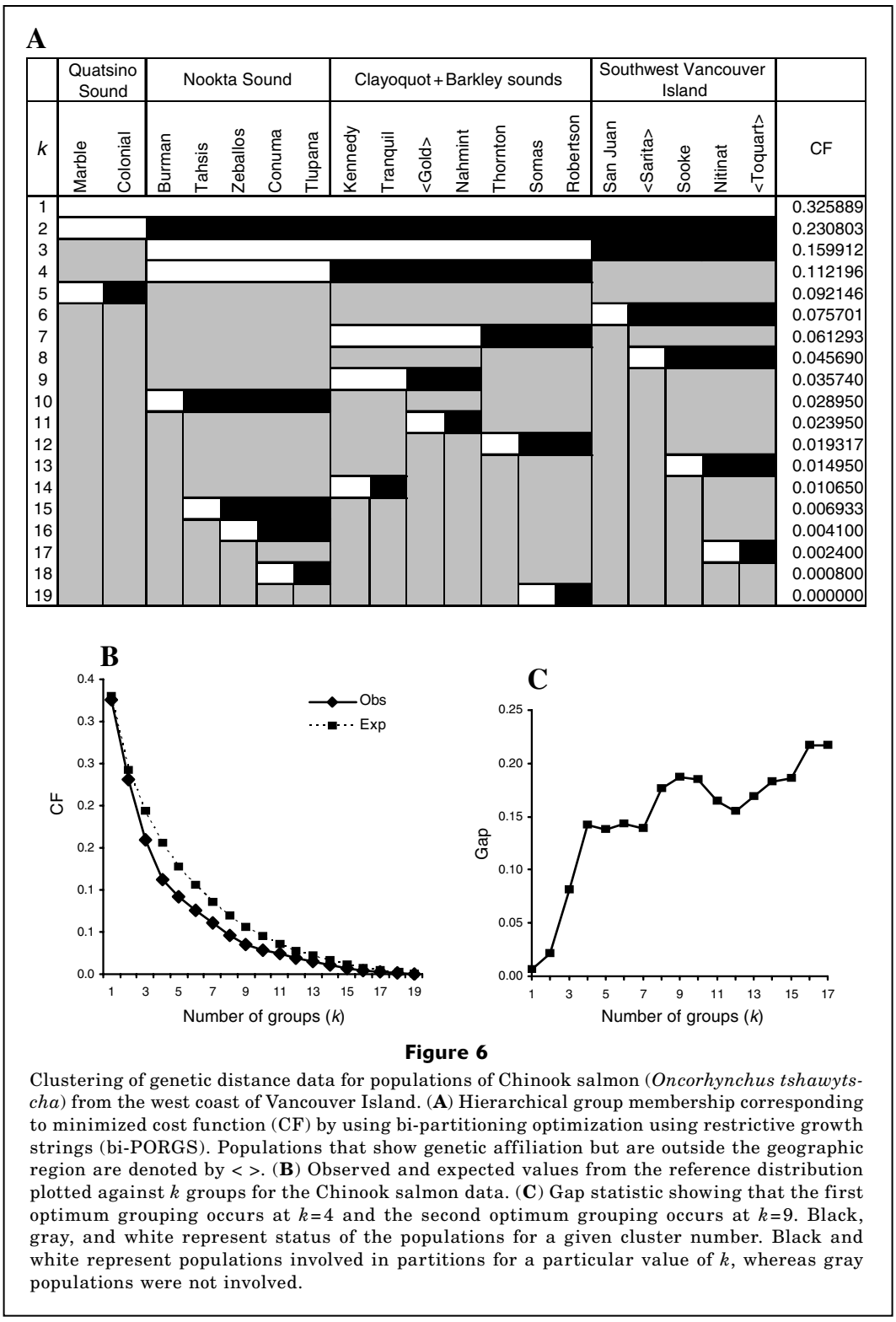


Figure 6

Clustering of genetic distance data for populations of Chinook salmon (*Oncorhynchus tshawytscha*) from the west coast of Vancouver Island. (**A**) Hierarchical group membership corresponding to minimized cost function (CF) by using bi-partitioning optimization using restrictive growth strings (bi-PORGS). Populations that show genetic affiliation but are outside the geographic region are denoted by < >. (**B**) Observed and expected values from the reference distribution plotted against k groups for the Chinook salmon data. (**C**) Gap statistic showing that the first optimum grouping occurs at k=4 and the second optimum grouping occurs at k=9. Black, gray, and white represent status of the populations for a given cluster number. Black and white represent populations involved in partitions for a particular value of k, whereas gray populations were not involved.

Weir and Cockerham's (1984) co-ancestory coefficient θ. A number of alternative distance measures could be tested with this method, but an examination of these measures is beyond the scope of this article. Also, determination and comparison of the optimal number of groupings directly from the multilocus genotypic data (e.g., Pritchard et al., 2000), instead of from the distance measures used here, would provide useful.

Unlike other clustering methods, PORGS does not have to embed distance data in vector space (i.e., mul-

tidimensional scaling); therefore, the underlying structure of the distance data remains intact, and resultant clusters can be compared between sets of populations. Successive increases in cluster numbers automatically lead to a hierarchical representation of the group structure. The gap statistic determines optimal number of groupings. In this example, k=4 was the first optimum number of groupings for Chinook salmon populations from the west coast of Vancouver Island. Except for populations impacted by straying fish and transferred fish, these groupings correspond to four geographic areas: Quatsino Sound, Nootka Sound, Clayoquot+Barkley sounds, and southwest Vancouver Island. Similar groupings were identified by agglomerative clustering seen in the UPGMA tree.

It was determined that random set partitions do not prevent visits to the same group memberships; therefore redundancy in cost function evaluations wastes processing time. Random search proves ineffective, except for very small n, because the prohibitively large number of iterations requires an unreasonable amount of time to find an optimal solution. However, an exhaustive search of the data space provided by a simple random search or a pass-through of all set partitions ensures that a globally optimized solution is found. By a globally optimal solution, we mean that no smaller cost function evaluations are possible for each k from a particular data set. Depending on available computational speed and the number of populations, PORGS can require an unreasonable amount of time. An alternative approach reduces the search by sequentially splitting into groups (bi-PORGS method) and evaluating subgrouping combinations to minimize the cost function. But like other hierarchical clustering methods, the nested search approach (bi-PORGS) means that prior cluster groups cannot be undone; therefore finding the optimal values may not always be possible. However, for the simulated data, PORGS and bi-PORGS methods produced the same results, indicating a globally optimal solution is possible with the nested search. The faster search method with bi-PORGS may forgo the guarantee of an optimal solution, but it can tackle larger problems, with the limitation being the number of populations in the first bipartition.

For large, coastwide data sets, a nested search requires bipartitioning a large number of populations simultaneously. Sparse data sets or optimization heuristics, such as thouse derived from deterministic annealing and mean field approximation, may be necessary when an exhaustive search is not possible (Puzicha et al., 1999). However, regional groupings could be recognized where each region could be run independently. This "divide and conquer" method requires that the subproblems be naturally disjoint, and that divisions be appropriate and of manageable size (Kirkpatrick et al., 1983). Ultimately, given the same set of genetic markers and distance measures, researchers will have a means of establishing groupings of varying size but representing similar levels of intracluster genetic variation.

Analysis of coded-wire tag data has indicated that straying Chinook salmon occur at a higher frequency between nearby spawning sites (e.g., Quinn, 1993; Candy and Beacham, 2000). Consequently, geographic distance between populations may be a good approximator of gene flow in salmon species; however, inferring barriers to migration on the basis of geographical or physical features alone can be misleading (Waples, 1991). The Gold River Chinook salmon population stands out by not conforming to the general rule of concordant genetic and geographic distance. According to geographic distance alone, Gold River Chinook salmon should be most genetically similar to Burman River Chinook salmon because less than 10 km separate the mouths of the two river systems. However, cluster analysis indicates that Gold River fish are most genetically similar to Barkley Sound fish, 125 km to the south (Fig. 1). Because the nearby Burman River population remains clustered with the Nootka Sound group, straying Barkley Sound fish must be extremely precise; apparently remaining in the Gold River only to spawn.

A number of factors could contribute to this restricted straying between Barkley Sound and the Gold River. Olfactory imprinting on waters near natal streams during out-migration is known to be important for successful homeward navigation (Harden Jones, 1968; Quinn, 1984). Consequently, the presence of pulp mills at the heads of both Muchalat (Gold River) and Alberni (Somas River) Inlets, and their effects on water chemistry, may increase straying between these two systems. Both systems lie at the head of long inlets, where the Gold and Somas Rivers have similar inlet and stream orientation. Also, both are lake-headed systems, possibly resulting in similarly modified river temperatures and flow regimes. Finally, approach to natal stream may be important for determining stray patterns. During the return migration to spawn, Barkley Sound, Chinook salmon heading south must first pass Nootka Sound, which provides an opportunity for these fish to eventually stray into the Gold River. The Gold River tissue samples collected in the early to mid-1980s, along with recent recoveries of thermally marked Robertson Hatchery fish in the Gold River, indicate that straying into the Gold River has likely occurred for quite a number of years.

Populations receiving transfers (Toquart, Thornton, and Sooke Rivers; Table 2) remain grouped to their respective donor stocks rather than to nearby populations, indicating that transfer history also plays an important role in establishing regional stock structure. The initial transfer of Robertson Creek fish to the Toquart River is not apparent from the bi-PORGS analysis, where Toquart River grouped with the second transfer source, Nitinat River. If native stocks existed in Toquart and Sooke Rivers before transfers into these systems, their continued existence there is not evident from the present study. However, populations with mixed ancestry may be better analyzed with individual-based clustering methods (Pritchard et al., 2000; Corander et al., 2003). The remaining two southwest Vancouver Island populations, where no transfers have occurred, remain quite distinctive.

Besides the history of transferred populations, other factors may determine genetic stock structure. Time of return to spawning grounds may provide a natural barrier to gene flow, preventing geographically superimposed populations from becoming genetically similar (Hendry and Day, 2005). Founder effects may play a role in shaping population structure, especially after recent colonization (Ramstad et al., 2004). Although multiyear sampling should address this problem, sampling error could be indistinguishable from allelic frequencies that are changed by some perturbing force. Indeed, small effective population size, where few related individuals are breeding, will hasten genetic drift (Waples, 1990). As a consequence of our inability to understand all mechanisms controlling gene flow, Waples (1991) warns against drawing inferences based on physical characteristics of the habitat without supporting biological information that links habitat differences to adaptations.

Little genetic variation with respect to population differentiation appears to have occurred in Robertson Creek over 23–30 years. Assuming that a majority of Robertson Creek fish return as four-year-olds (Healey, 1991), these years represent six to eight generations of Chinook salmon. The stability of microsatellite markers has been reported elsewhere for Atlantic salmon (*Salmo salar*) over a time frame of three to five generations (Tessier and Bernatchez, 1999). Furthermore, the genetic variation between populations with microsatellite markers was found to be 19 times greater than the interannual variation for sockeye salmon (*Oncorhynchus nerka;* Beacham et al., 2006b).

Microsatellites provide highly stable, reliable genetic markers for comparisons of genetic variation across the range of a species and are thus becoming an important tool for the management and conservation of genetic diversity of Pacific salmon species. Although genetic characters detected with these markers are neutral with respect to natural selection, it is likely that they are indicators of local adaptation in other encoding parts of the genome (Waples, 1991). Fine-scale grouping of genetically similar populations allows managers to make informed harvest and enhancement decisions. As was evident with Chinook salmon from the west coast of Vancouver Island, strictly geographically based assumptions regarding the level of genetic relatedness between populations can be incorrect.

Acknowledgements

The computer program PORGS, using the cost function (Roth et al., 2003) and RGS (Knuth, 2005; Algorithm 7.2.1.5H, modified for *r* blocks) is available from the contact author or for downloading from http://www.pac.dfo-mpo.gc.ca/sci/mgl/data_e.htm. We thank the staff at Robertson, Nitinat, and Conuma hatcheries, and R. Dunlop of the Nuu-chah-nulth Tribal Fisheries, for providing tissue samples for this analysis. We thank staff of the Molecular Genetics Laboratory (M. Wetklo, K. Jonsen, and J. Supernault) for laboratory work. We also thank the contributions of three anonymous reviewers who helped provide focus and clarity to the methods portion of this article.

Literature cited

Beacham, T. D., K. L. Jonsen, J. Supernault, M. Wetklo, L. Deng, and N. Varnavskaya.
2006a. Pacific Rim population structure of Chinook salmon as determined from microsatellite variation. Trans. Am. Fish. Soc. 135:1604–1621.

Beacham, T. D., B. McIntosh, C. MacConnachie, K. M. Miller, R. E. Withler, and N. Varnavskaya.
2006b. Pacific Rim population structure of sockeye salmon as determined from microsatellite analysis. Trans. Am. Fish. Soc. 135:174–187.

Buhmann, J. M.
2002. Data clustering and learning. *In* Handbook of brain theory and neural networks, 2nd ed. (M. A. Arbib, ed.), p. 308–312. MIT Press, Cambridge, MA.

Cameron, P. J.
1994. Combinatronics, Topics, Techniques, and Algorithms, 355 p. Cambridge Univ. Press, Cambridge, UK.

Candy, J. R., and T. D. Beacham.
2000. Patterns of homing and straying in southern British Columbia coded-wire tagged chinook salmon (*Oncorhynchus tshawytscha*) populations. Fish. Res. 4:41–56.

Cavalli-Sforza, L. L., and A. W. F. Edwards.
1967. Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570.

Corander, J., P. Waldmann, and M. J. Sillanpää.
2003. Bayesian analysis of genetic differentiation between populations. Genetics 163:367–374.

Cross, C. L., L. Lapi, and E. A. Perry.
1991. Production of Chinook and Coho salmon from British Columbia hatcheries, 1971 through 1989. Can. Tech. Rep. Fish. Aquat. Sci. 1816, 48 p.

Felsenstein, J.
1985. Confidence limits on phylogenies: an approach using the bootstrap. Evol. 39:783–791.
1989. PHYLIP—Phylogeny inference package, vers. 3.2. Cladistics 5**:**164–166.

Goudet, J.
1995. FSTAT, a program to calculate *F*-statistics, vers, 1.2. J. Hered. 86:485–486.

Harden Jones, F. R.
1968. Fish migration, 134 p. St. Martin's Press, New York, NY.

Healey, M. C.
1991. Life history of Chinook salmon (*Oncorhynchus tshawytscha*). *In* Pacific salmon life history (C. Groot, and L. Margolis, eds.), p. 313–393. Univ. British Columbia Press, Vancouver, B.C.

Hendry, A. P., and T. Day.
2005. Population structure attributable to reproductive time: isolation by time and adaptation by time. Mol. Ecol. 14:901–916.

Hofmann, T., and J. M. Buhmann.
1997. Pairwise data clustering by deterministic annealing. IEEE Trans. Pattern Anal. Mach. Intel. 18:1–37.

Jain, A. K., M. N. Murty, and P. J. Flynn.
1999. Data clustering: a review. Assoc. Comput. Mach. Trans. Comput. Surv.31:265–322.
Kirkpatrick, S., C. D. Gelitt, and M. P. Vecchi.
1983. Optimization by simulated annealing. Science 220:671–680.
Knuth, D. E.
2005. The art of computer programming, vol. 4, fascicle 3, Generating all combinations and partitions, 150 p. Addison-Wesley, Reading, MA.
Nei, M.
1987. Molecular evolutionary genetics, 512 p. Columbia Univ. Press, New York, NY.
Nei, M., F. Tajima, and Y. Tateno.
1983. Accuracy of estimated phylogenetic trees from microsatellite DNA. Genetics 144:389–399.
Pritchard, J. K., M. Stephens, and P. Donnelly.
2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Puzicha, J., T. Hofmann, and J. Buhmann.
1999. A theory of proximity based clustering: structure detection by optimization. Pattern Recogn. 33:617–634.
Quinn, T. P.
1984. Homing and straying in Chinook salmon (*Oncorhynchus tshawytscha*) from the Cowlitz River hatchery, Washington. Can. J. Fish. Aquat. Sci. 41:1078–1082.
1993. A review of homing and straying of wild and hatchery-produced salmon. Fish. Res. 18:29–44.
Quinn, T. P., and A. H. Dittman.
1990. Pacific salmon migrations and homing: mechanisms and adaptive significance. Trends Ecol. Evol. 5:174–177.
Ramstad, K. M., C. A. Woody, G. K. Sage, and F. W. Allendorf.
2004. Founding events influence genetic population structure of sockeye salmon (*Oncorhynchus nerka*). Mol. Ecol. 13:277–290.
Ricker, W. E.
1972. Hereditary and environmental factors affecting certain salmonid populations. *In* The stock concept of Pacific salmon (R. C. Simon, and P. Larkin, eds.), p. 19–160. Univ. British Columbia Press, Vancouver, B.C.
Riddell, B. E.
1993. Spatial organization of Pacific salmon: what to conserve? *In* Genetic conservation of salmonid fishes (J. G. Cloud, and G. H. Thorgaard, eds.), p. 23–41. Plenum Press, New York, NY.
Robinson, W. S.
1951. A method for chronologically ordering archeological deposits. Am. Antiq. 16:293–301.
Rota, G. C.
1964. The number of partitions of a set. Am. Math. Mon. 71(5):498–504.
Roth, V., J. Laub, M. Kawanabe, and J. Buhmann.
2003. Optimal cluster preserving embedding of nonmetric proximity data. IEEE Trans. Pattern Anal. Mach. Intel. 5:1540–1550.
Sneath, P. H. A., and R. R. Sokal.
1973. Numerical taxonomy, 573 p. Freeman, San Francisco, CA.
Swofford, D. L., G. L. Olsen, P. J. Waddell, and D. M. Hillis.
1996. Phylogenetic inference. *In* Molecular systematics (D. M. Hillis and C. Mortiz, eds.), p. 407–514. Sinauer Assocs., Dunderland, MA.
Teel, D. J., G. B. Miller, G. A. Winans, and W. S. Grant.
2000. Genetic population structure and origin of life history types in Chinook salmon in British Columbia, Canada. Trans. Am. Fish. Soc. 129:194–209.
Tessier, N., and L. Bernatchez.
1999. Stability of population structure and genetic diversity across generations assessed by microsatellites among sympatric populations of landlocked Atlantic salmon (*Salmo salar* L.). Mol. Ecol. 8:169–179.
Tibshirani, R., G. Walther, and T. Hastie.
2001. Estimating the number of clusters in a data set via the Gap statistic. J. Royal Stat. Soc. 63(2):411–423.
van der Kloot, W. A., A. M. J. Spaans, and W. J. Heiser.
2005. Instability of hierarchical cluster analysis due to input order of the data: the Permcluster solution. Psychol. Methods 10:468–476.
Waples, R. S.
1990. Conservation genetics of Pacific salmon. II. Effective population size and the rate of loss of genetic variability. J. Hered. 81:256–276.
1991. Pacific salmon, *Onchorhynchus* spp., and the definition of "species" under the Endangered Species Act. Mar. Fish. Rev 53:11–22.
Waples, R. S., and O. Gaggiotti.
2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol. Ecol. 15:1419–1439.
Waples, R. S., R. G. Gustafson, L. A. Weitkamp, J. M. Myers, O. W. Johnson, P. J. Busby, J. J. Hard, G. J. Bryant, F. W. Waknitz, K. Nelly, D. Teel, W. S. Grant, G. A. Winans, S. Phelps, A. Marshall, and B. M. Baker.
2001. Characterizing diversity in salmon from the Pacific Northwest. J. Fish. Biol. 59:1–41.
Weir, B. S., and C. C. Cockerham.
1984. Estimating *F*-statistics from the analysis of population structure. Evolution 38:1358–1370.