NPAFC
Doc. 908
Rev. _____

SNPs provide high-throughput resolution for migratory studies of Chinook salmon

by

William D. Templin, Christian T. Smith, James E. Seeb, and Lisa W. Seeb

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

Submitted to the

NORTH PACIFIC ANADROMOUS FISH COMMISSION

BY THE

UNITED STATES OF AMERICA

October, 2005

THIS PAPER MAY BE CITED IN THE FOLLOWING FORMAT:
Templin, W. D., C. T. Smith, J. E. Seeb, and L. W. Seeb. 2005. SNPs provide high-throughput resolution for migratory studies of Chinook salmon. (NPAFC Doc. 908) 10 p. Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, AK, USA 99518.

Abstract

Genetic analyses of salmon in international waters require large, comprehensive datasets of genetic information developed cooperatively by multiple parties across international boundaries. Multi-party, multi-national datasets require robust methods that are both transparent and transportable because all parties are contributing to a common set of data that can be accessed and used by each for analyses. We developed 33 markers based on single nucleotide polymorphisms (SNPs) to rapidly genotype large numbers of individuals for mixed stock identification of local Alaska fishery harvests. SNP genotypes can be unambiguously assayed by a wide array of techniques under many different conditions, making markers of this type ideal for multi-national, multi-laboratory studies of shared salmon resources. The transparency and transportability of these data has been demonstrated both regionally and internationally. We compare SNPs from representative populations originating from throughout the range of Chinook salmon to determine their utility for mixed stock analysis of high seas samples. Our results to date show that SNP markers are a rapid and cost effective approach to analysis of the large number of samples from complex mixtures encountered in multi-national research and fishery monitoring efforts.

Background

In 2000, the North Pacific Anadromous Fish Commission established an *ad hoc* working group on stock identification with the intent to "develop, standardize, and disseminate genetic and other databases among the Parties, to encourage the development of new genetic technologies, and to facilitate the dissemination of statistical techniques." Genetic analyses of salmon in international waters require large, comprehensive datasets of genetic information developed cooperatively by multiple parties across international boundaries. Multi-party, multi-national datasets require robust methods that are both transparent and transportable because all parties are contributing to a common set of data that can be accessed and used by each for analyses. The marker classes most commonly proposed to meet these requirements have been allozymes, microsatellites, and, more recently, single nucleotide polymorphisms (SNPs). Allozymes have provided valuable insights over several decades and a fully standardized baseline is currently available for mixed stock analyses of Chinook salmon (*Oncorhynchus tshawytscha*) (Teel et al. 1999), but its use has subsided due to the high throughput rates and improved resolution available via DNA markers (microsatellites and SNPs).

Microsatellite loci in Chinook salmon have proven very powerful in population structure and mixture analyses (Banks et al. 2000; Nelson et al. 2001). A limitation of microsatellite data is that data generated in one laboratory are not readily combined with data collected in another. This is because raw data are dependent on specific aspects of hardware and chemistry installed in a given laboratory (Wattier et al. 1998; Haberl & Tautz 1999) and on the environment in the laboratory in which data are collected (Davison & Chiba 2003).

For Chinook salmon the cost and effort required to standardize microsatellites has been accomplished through an eight-laboratory collaboration funded by the Chinook Technical

Committee (CTC) of the Pacific Salmon Commission. Over the course of two years and at the expense of US$1.1 million, 62 candidate microsatellite loci were evaluated; 13 loci were chosen for use based on their robustness and consistency under multiple laboratory conditions. From this effort a baseline of 110 populations from Southeast Alaska, Canada and the Pacific Northwest United States has been created. This baseline will be used in the near future to estimate stock composition in fisheries covered under the Pacific Salmon Treaty. Other Pacific Salmon species do not currently have a similar dataset that involves more than two or three laboratories.

Unlike microsatellites, SNPs are standardized by definition and do not require extensive and costly standardization efforts. They are an efficient method for achieving the requirements of common-use datasets. SNPs are generally biallelic, are distributed throughout the genome, and the laboratory analysis and scoring of raw data are highly amenable to automation. Because the measurement of a SNP involves measuring a primary characteristic (nucleotide type; A, C, G, or T) and not a secondary characteristic (e.g. fragment length for microsatellites), the genotype information is unambiguous (not effected by the means of analysis). The widespread nature of SNPs in the genome enables analyses to include many previously inaccessible areas of the DNA, for example the inclusion of information from regions under selection. Finally, these markers can be analyzed in an efficient and cost effective manner. For example, a single technician at the Alaska Department of Fish and Game (ADF&G) recently assayed (lab analysis, scoring, and data entry) more than 1800 individuals at 33 SNP loci in 10 workdays at a normal pace. After quality control checks, the error rate was determined to be less than 0.01% and the failure rate was 1.6%.

SNPs have also been demonstrated to be easily transportable and repeatable. We have successfully exchanged, checked and merged data with five other laboratories (Hokkaido University, VNIRO, Columbia River Intertribal Fisheries Commission, Washington Department of Fisheries, NOAA-Auke Bay Laboratory, and NOAA-Montlake Laboratory) conducting baseline studies on Chinook, chum (*O. keta*) and sockeye salmon (*O. nerka*).

The Pacific Salmon Commission has decided that future additions of markers to the Chinook salmon baseline will be SNPs, and they have already funded laboratories to continue the discovery and evaluation of SNP markers in Chinook salmon.

Alaska SNP Baseline Development

Recognizing its position in the center of Pacific salmon production, in both geography and abundance, the State of Alaska has actively pursued the development of datasets that are useful on an international scale. During the 1990's genetic data from Alaska populations of Chinook chum, and sockeye salmon to standardized datasets (Teel et al. 1999; Habicht et al. 2001; Seeb et al. 2004) of allozyme loci for use with marine mixtures. When technology developed sufficiently to give DNA-based techniques the advantage over allozymes, numerous techniques were developed and evaluated for the eventual replacement of the allozyme baseline (e.g. Bentzen et al. 1991; Allendorf and Seeb 2000; Banks et al. 2000). While many of the new techniques were powerful for identifying fine-scale differences between salmon, most suffered from issues such as slow throughput or non-transportability. For this reason ADF&G began to

investigate the utility of SNPs for development of large-scale multi-national baselines for analysis of marine mixtures.

Currently, ADF&G has identified 33 SNPs (10 reported in Smith et al. 2005a) in Chinook salmon and most are being surveyed in the major Chinook salmon-producing drainages of Alaska. The most complete set of data exists for the Yukon River, where 23 populations were originally surveyed for variation at 10 SNPs. Sufficient variation was found within these markers to exceed the requirements for identification of management units and meet the obligations of the Pacific Salmon Treaty between the U.S. and Canada (Smith et al. 2005b). This baseline has been augmented to 26 SNPs and was used by ADF&G to estimate the stock composition of U.S. fisheries in the Yukon River in 2004. Within the Kuskokwim River and Copper River drainages, 15 populations each are being surveyed for variation at 33 SNPs with plans to use these data for analysis of stock composition in fishery harvests in nearby marine waters. The ADF&G Gene Conservation Laboratory will soon begin analysis of the full set of SNP loci within Norton Sound (five populations), Cook Inlet (15 populations) and Southeast Alaska (~10 populations). Analysis of the Southeast Alaska populations will be complemented by a Pacific Salmon Commission-funded survey of SNPs in 35 populations of Chinook salmon from Southeast Alaska, British Columbia, Washington, Oregon, Idaho, and California. In addition, at least four populations from the Kamchatka Peninsula are available for inclusion to increase the geographic range of present analyses.

Large-scale Comparison

Here, we present an initial analysis of the utility of SNPs for stock identification applications in mixtures of interest to the NPAFC. We selected 39 populations of Chinook salmon from across the species range as an initial representation of potential variation at a common set of 25 SNPs (Table 1; Figure 1). All 23 populations from the Yukon River were included in the analysis as a measure of the scale of genetic variation. Previous analyses have demonstrated that the populations from western Alaska showed little variation at allozyme loci across a wide geographic range (Gharrett et al. 1987). The presence of variation among collections in this region would suggest the potential value of a range-wide baseline of SNPs.

The degree of genetic similarity between populations was measured by computing inter-population genetic distances (Cavalli-Sforza and Edwards 1967). The pattern of genetic similarity was visualized by clustering populations based on these genetic distances using an averaging algorithm (UPGMA; Sneath and Sokal 1973) to create a tree (Figure 2), where population location within groups is presented by a branching network. Branch distance between any two populations is an indication of their genetic similarity, with shorter branches suggesting greater similarity. From this tree it is apparent that major differences within Chinook salmon are detected by these 25 SNPs. The most divergent population in this dataset is an ocean-type population from the Columbia River, followed by a Russian population and two stream-type populations from the Columbia River. In agreement with previous studies (Gharrett et al. 1987), there is lower, but still significant, variation between populations within Alaska. The tree shows evidence of genetic variation that generally corresponds to geography.

We defined groups of populations for identification in mixtures based on the genetic similarity observed between these populations (Figure 2). Due to the consistency with which genetic similarity matches geographic proximity, these groups define regions closely matching the geographic and political boundaries of the Chinook salmon freshwater habitat.

The utility of these reporting groups for genetic stock identification of mixtures of Chinook salmon taken from national and international marine waters was evaluated through simulations. These simulations were designed to assess whether the baseline of SNP allele frequencies provides sufficient information to identify reporting groups (stocks) in hypothetical mixtures. Simulations were performed using the Statistical Package for Analyzing Mixtures (SPAM version 3.6, Debevec et al. 2000) to estimate the composition of a hypothetical mixture of predetermined stock proportions. This process involved an iterative process during which the mixture genotypes and baseline frequencies were randomly generated from the known baseline allele frequencies assuming Hardy-Weinberg equilibrium. Mean estimates of mixture proportions and 90% confidence intervals were derived from the results of 1000 iterations. The lower and upper bounds of the confidence intervals were determined by sorting the estimates and selecting the 51^{st} and 950^{th} results. Mixtures simulated (N = 400) were entirely composed (100%) of a single reporting group; repeated for each reporting group. When a reporting group mixture was simulated, all baseline populations in the group contributed equally to the mixture. Reporting groups with mean correct estimates of 90% or better are considered highly identifiable in potential mixtures. Reporting groups with mean correct estimates lower than 90% can still be considered identifiable in mixtures, but sources of misallocation should be considered when interpreting the results. The results of these simulations demonstrate a degree of precision and accuracy that is useful for identification of stock composition (Table 1). The two groups with the lowest accuracy (Kuskokwim and Togiak) were just below 90%. Achieving this level of accuracy is encouraging as western Alaska populations are chronically the most difficult to distinguish.

Conclusions

While the data set presented here cannot be construed to represent the range of variety found within Chinook salmon populations of the North Pacific, it does provide a useful demonstration of the power of a large-scale SNP baseline for mixtures of interest to the NPAFC.

Multi-party, multi-national datasets require robust methods that are both transparent and transportable because all parties are contributing to a common set of data that can be accessed and used by each for analyses. SNPs offer several important advantages for cooperative studies among international laboratories: 1) rapid and unambiguous scoring, 2) data are not lab-specific and may be readily shared, 3) new chemistries offer very high-throughput (~10-fold of that typical of other DNA types), and 4) the cost per fish is decreasing. We are currently cooperating with 10 other laboratories on SNP development and applications in Chinook salmon. We are readily exchanging sockeye SNP data with VINRO and chum salmon SNP data with Hokkaido University. Six U.S. laboratories and one Russian laboratory have installed modern high-throughput SNP detection hardware and two Japanese laboratories score SNPs by microarray. For these reasons, we believe that SNPS will be a major part of NPAFC studies in the future.

Literature Cited

Allendorf, F. W. and L. W. Seeb. 2000. Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. Evolution 54:640-651.

Banks M. A., V. K. Rashbrook, M. J. Calavetta, C. A. Dean, D. Hedgecock. 2000. Analysis of microsatellite DNA resolves genetic structure and diversity of chinook salmon (Oncorhynchus tshawytscha) in California's Central Valley. Canadian Journal of Fisheries and Aquatic Sciences 57:915-927.

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

Davison A. and S. Chiba. 2003. Laboratory Temperature Variation Is a Previously Unrecognized Source of Genotyping Error During Capillary Electrophoresis. Molecular Ecology Notes 3:321-323.

Debevec, E. M., R. B. Gates, M. Masuda, J. Pella, J. Reynolds, and L. W. Seeb. 2000. SPAM (Version 3.2): Statistics program for analyzing mixtures. Journal of Heredity 91:509-511.

Gharrett, A. J., S. M. Shirley and G. R. Tromble. 1987. Genetic relationships among populations of Alaskan chinook salmon (Oncorhynchus tshawytscha). Canadian Journal of Fisheries and Aquatic Sciences 44:765-774.

Habicht, C, W.D Templin, C. M Guthrie III, R. L. Wilmot, G.A. Winans, E. Iwamoto, J.E. Seeb and L.W. Seeb. 2001. Status report for genetic stock identification studies of Pacific Rim sockeye salmon. (NPAFC Doc. 562). Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, Alaska 99802.

Haberl M., D. Tautz. 1999. Comparative allele sizing can produce inaccurate allele size differences for microsatellites. Molecular Ecology 8:1347-9.

Nelson, R. J., M. P. Small, T. D. Beacham, and K. J. Supernault. 2001. Population structure of Fraser River Chinook salmon (Oncorhynchus tshawytscha): an analysis using microsatellite DNA markers. Fishery Bulletin 99: 94-107.

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

Smith, C. T., J.E. Seeb, P. Schwenke, and L.W. Seeb. 2005a. Use of the 5'-nuclease reaction for SNP genotyping in Chinook salmon. Transactions of the American Fisheries Society 134:204-217

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

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. W. H. Freeman, San Francisco, California, 573 pp.

Teel, D. J., P. A. Crane, C. M. Guthrie III, A. R. Marshall, D. M. Van Doornik, W. D. Templin, N. V. Varnavskaya, and L. W. Seeb. 1999. Comprehensive allozyme database discriminates chinook salmon around the Pacific Rim (NPAFC document 440) 25p. Alaska Department of Fish and Game, Division of Commercial Fisheries, 333 Raspberry Road, Anchorage, Alaska USA 99518.

Wattier, R., C.R. Engel, P. Saumitou-Laprade, M. Valero, 1998. Short Allele Dominance as a Source of Heterozygote Deficiency at Microsatellite Loci: Experimental Evidence at the Dinucleotide Locus Gv1ct in Gracilaria Gracilis (Rhodophyta). Molecular Ecology 7:1569-1573.

Table 1. Regional groups of Chinook salmon populations used to investigate the potential utility of a SNP baseline for analysis of mixtures of interest to the NPAFC. Correct mean allocation and 90% confidence intervals from 100% regional simulations are indicated. Region numbers refer to locations in Figure 1.

Region		Number of	Simulation Results	
#	Name	Populations	Mean	90% CI
1	Russia	1	0.986	(0.947-1.000)
2	Lower Yukon	4	0.955	(0.904-0.988)
3	Middle Yukon	6	0.974	(0.944-0.996)
4	Canadian Yukon	13	0.986	(0.962-1.000)
5	Kuskokwim	1	0.899	(0.817-0.976)
6	Togiak	1	0.898	(0.808-0.970)
7	Bristol Bay	1	0.973	(0.938-0.996)
8	Kodiak	1	0.980	(0.935-0.998)
9	Susitna	1	0.921	(0.783-1.000)
10	Kenai	1	0.953	(0.885-0.993)
11	Chilkat	2	0.976	(0.930-0.997)
12	S. Southeast Alaska	3	0.990	(0.976-0.999)
13	King Salmon	1	0.972	(0.928-1.000)
14	Columbia ocean-type	1	0.984	(0.952-1.000)
15	Columbia stream-type	2	0.984	(0.954-1.000)

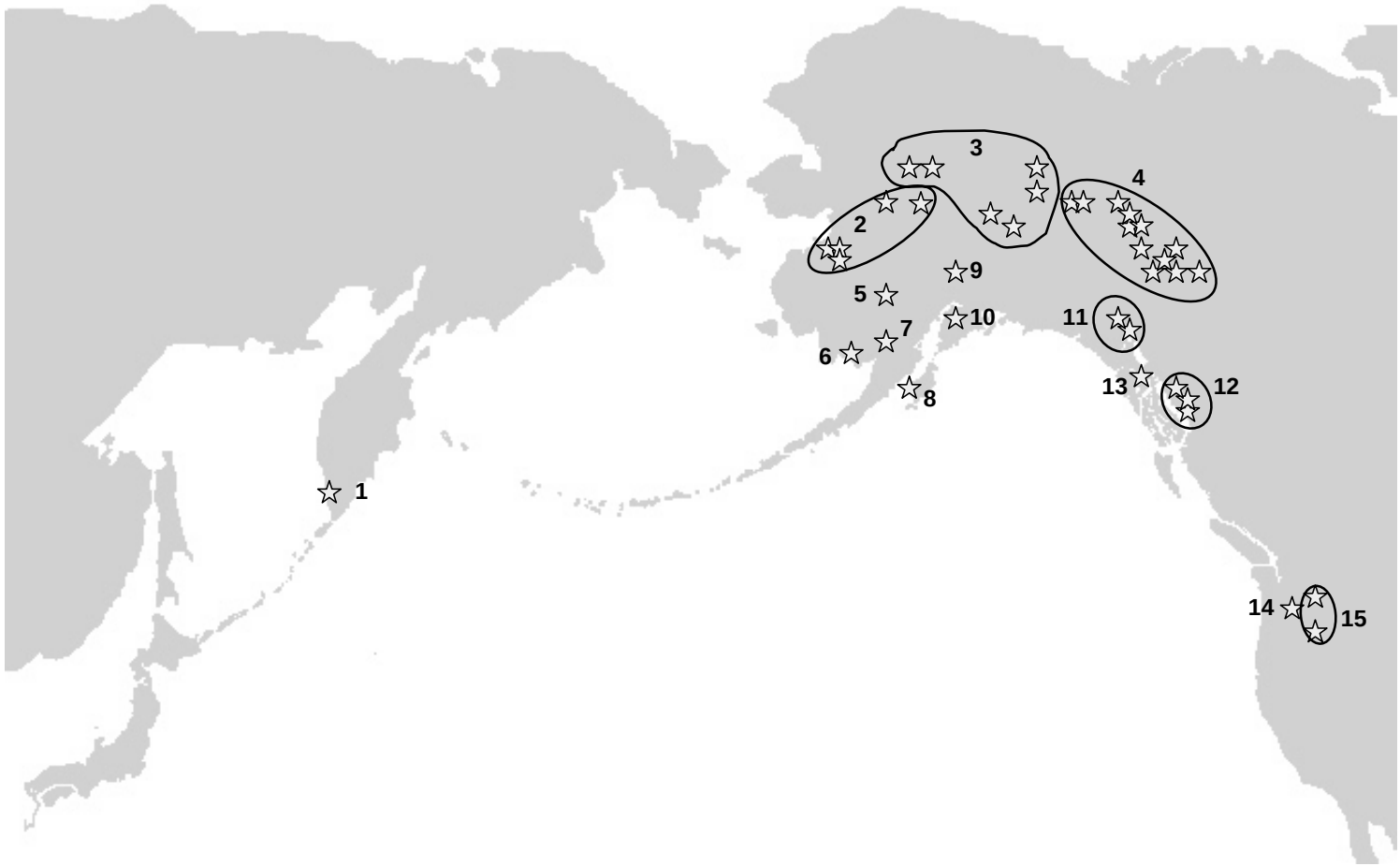

Figure 1. Locations (star symbols) of the 39 Chinook salmon populations used in the analysis of SNP loci on a Pacific Rim scale. Numbers refer to regional groups identified in Table 1.

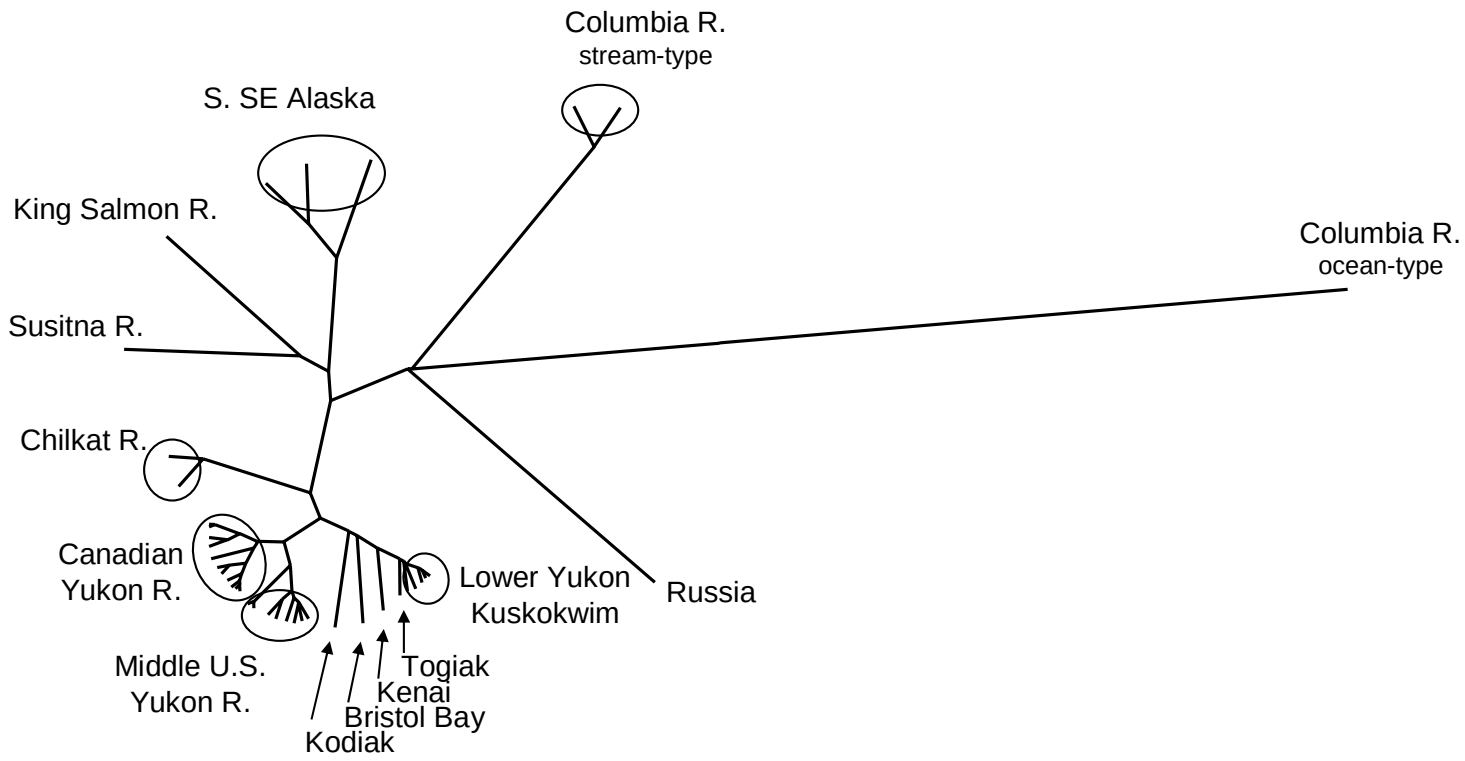

Figure 2. Unrooted tree created by the unweighted pair group mean algorithm (UPGMA) of Cavalli-Sforza and Edwards chord distances calculated between population pairs; ovals indicate regional groups.