Finding the best subset of SNPs for distinguishing populations of Yukon River Chinook salmon

William D. Templin and Anton B. Antonovich

Division of Commercial Fisheries, Alaska Department of Fish and Game, Anchorage, Alaska, USA

Introduction

Recent advances in laboratory methods have increased the number of genetic markers available for identifying stock components in mixtures. Although the cost and laboratory time to analyze each marker has decreased, overall costs have increased due to running ever larger numbers of markers. Prior to the availability of a large number of markers most researchers analyzed all available markers to maximize precision. Increasingly, selecting the most informative markers for specific applications will be critical to containing costs while maximizing the discrimination of key stock components. Here we explored three methods of choosing an optimal set of single nucleotide polymorphism (SNP) loci from the 24 SNPs available in Yukon River Chinook salmon populations. By ranking each locus using 1) mean interpopulation allelic frequency differences (*delta*), 2) mean interpopulation Fst, and 3) summed loadings for each locus from Principal Components Analysis (PCA), we developed sets of informative loci that were incrementally tested for precision and accuracy in simulated mixed stock analysis.

Methods

Ranking Markers -

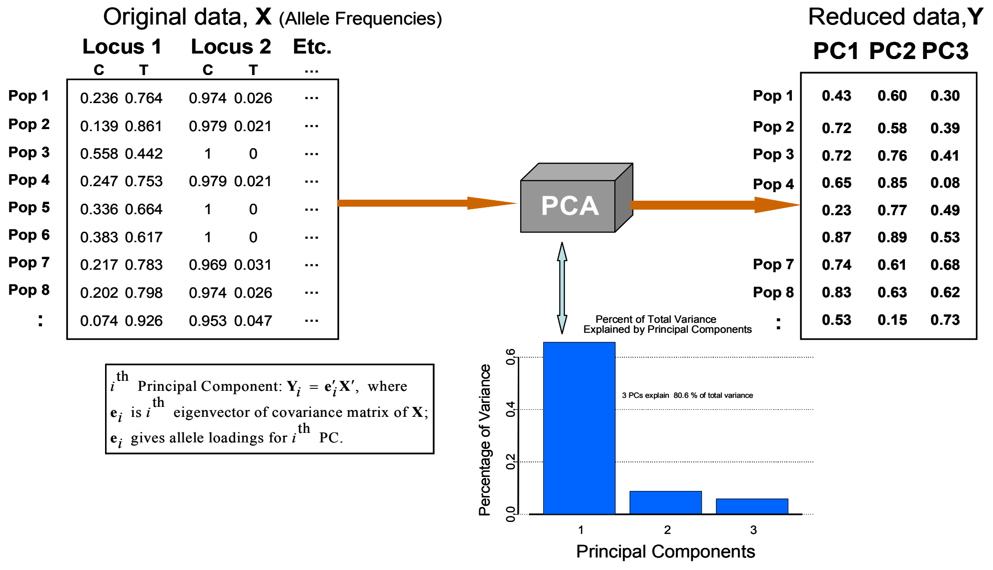
delta - The *delta* statistic measures the genetic distance between population pairs as the sum of the absolute differences between allele frequencies. Markers were ranked by the mean of the interpopulation *delta* values calculated for that marker.

$$\delta_{AB} = \frac{1}{2}\sum_{i=1}^{L}\left|p_i^A - p_i^B\right|$$

Fst - The Fst statistic is a measure of genetic diversity based on partitioning the variance of allele frequencies within and among populations in a "weighted" ANOVA. Markers were ranked by the mean of the interpopulation Fst values calculated for that marker.

PCA (Principal Components Analysis) - PCA is used as a data reduction method that seeks to explain the variation in data (allele frequencies) with fewer parameters. We adapt this method to provide information about which markers are more closely associated with the variation in allele frequency within the data set. A brief description of the method follows:

Part 1: Find Principal Components that account for >80% of variation.



Original data, **X** (Allele Frequencies)

	Locus 1 C	Locus 1 T	Locus 2 C	Locus 2 T	Etc. ...
Pop 1	0.236	0.764	0.974	0.026	...
Pop 2	0.139	0.861	0.979	0.021	...
Pop 3	0.558	0.442	1	0	...
Pop 4	0.247	0.753	0.979	0.021	...
Pop 5	0.336	0.664	1	0	...
Pop 6	0.383	0.617	1	0	...
Pop 7	0.217	0.783	0.969	0.031	...
Pop 8	0.202	0.798	0.974	0.026	...
:	0.074	0.926	0.953	0.047	...

Reduced data, **Y**

	PC1	PC2	PC3
Pop 1	0.43	0.60	0.30
Pop 2	0.72	0.58	0.39
Pop 3	0.72	0.76	0.41
Pop 4	0.65	0.85	0.08
	0.23	0.77	0.49
	0.87	0.89	0.53
Pop 7	0.74	0.61	0.68
Pop 8	0.83	0.63	0.62
:	0.53	0.15	0.73

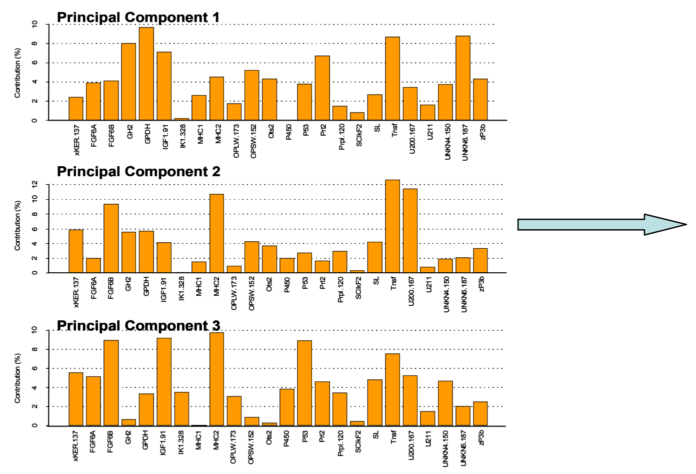
i^{th} Principal Component: $\mathbf{Y}_i = \mathbf{e}_i'\mathbf{X}'$, where
$\mathbf{e}_i$ is i^{th} eigenvector of covariance matrix of $\mathbf{X}$;
$\mathbf{e}_i$ gives allele loadings for i^{th} PC.

Percent of Total Variance
Explained by Principal Components
3 PCs explain 80.6 % of total variance
Percentage of Variance
0.0
0.2
0.4
0.6
1
2
3
Principal Components

Part 2: Determine each marker contribution to each Principal Component and rank by average.

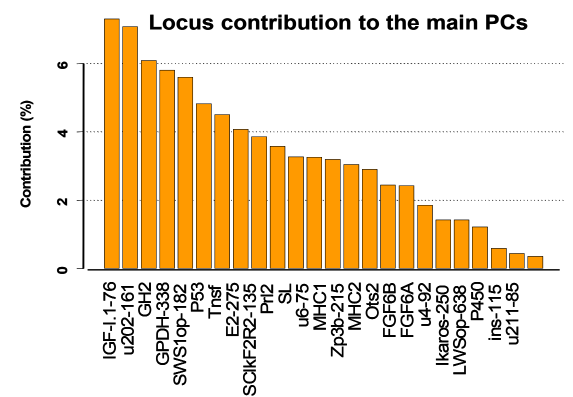
- Each allele gets a loading for each PC
- Sum loadings over alleles w/in a locus for locus contribution to a PC
- Loci with high contributions to main PCs are the best

- Average locus contributions over the PCs weighted by the percent of total variance explained by PC (i.e. importance)
- Rank loci by highest weighted average contribution.



$$Contr_i^L = \sum_{j=1:nAll}\left|e_{ij}^L\right|, \text{ where}$$

e_{ij} is a loading for j^{th} allele of locus L in i^{th} PC.



$$Rank^L = \sum_{i=1:nPCs} w_i * \sum_{j=1:nAll}\left|e_{ij}^L\right|, \text{ where}$$

$w_i = \dfrac{\lambda_i}{\sum \lambda_i}$ is the weight of i^{th} PC,

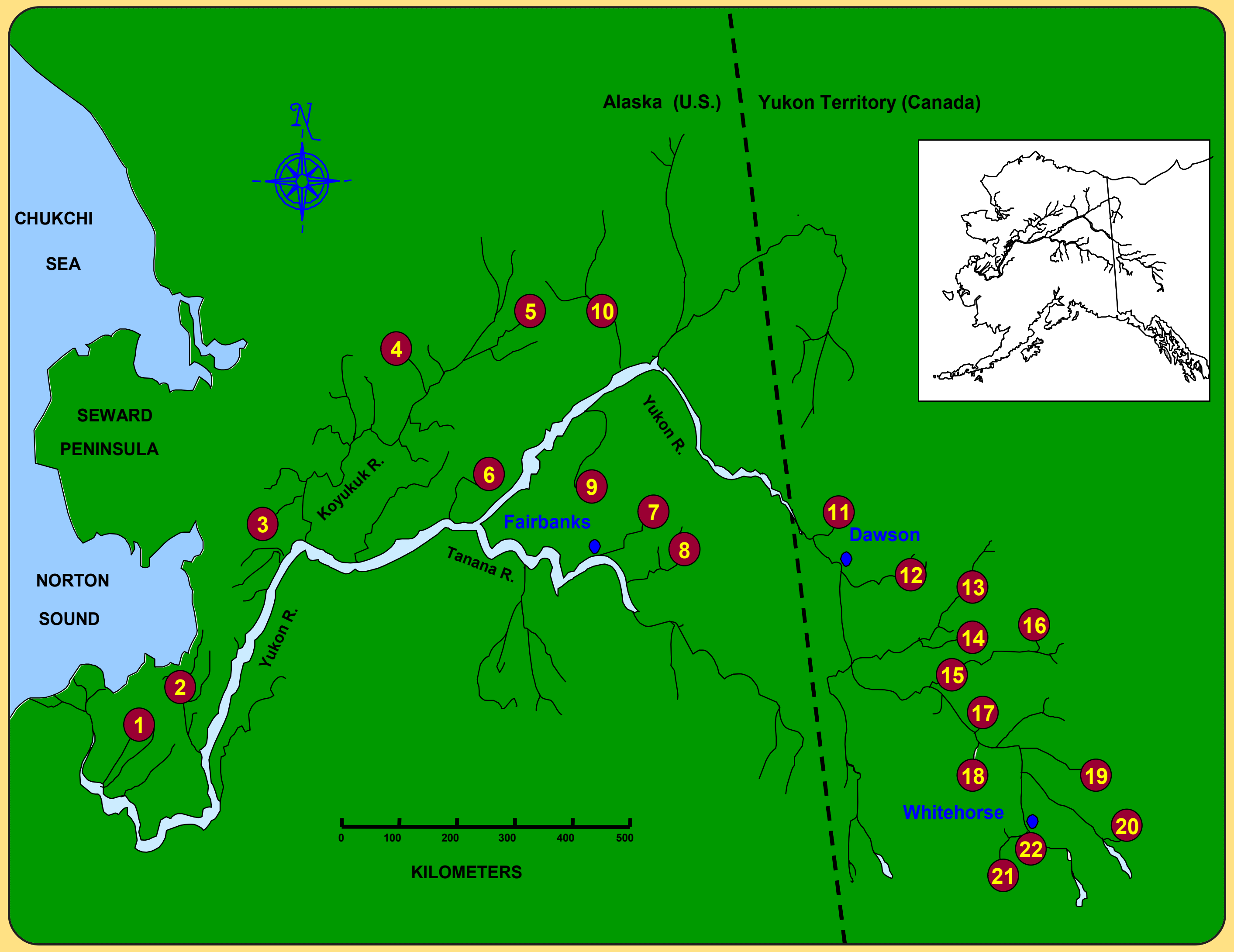
λ_i is i^{th} eigenvalue of Cov($\mathbf{X}$) and

e_{ij} is a loading for j^{th} allele of locus L in i^{th} PC.

Testing Marker Sets -

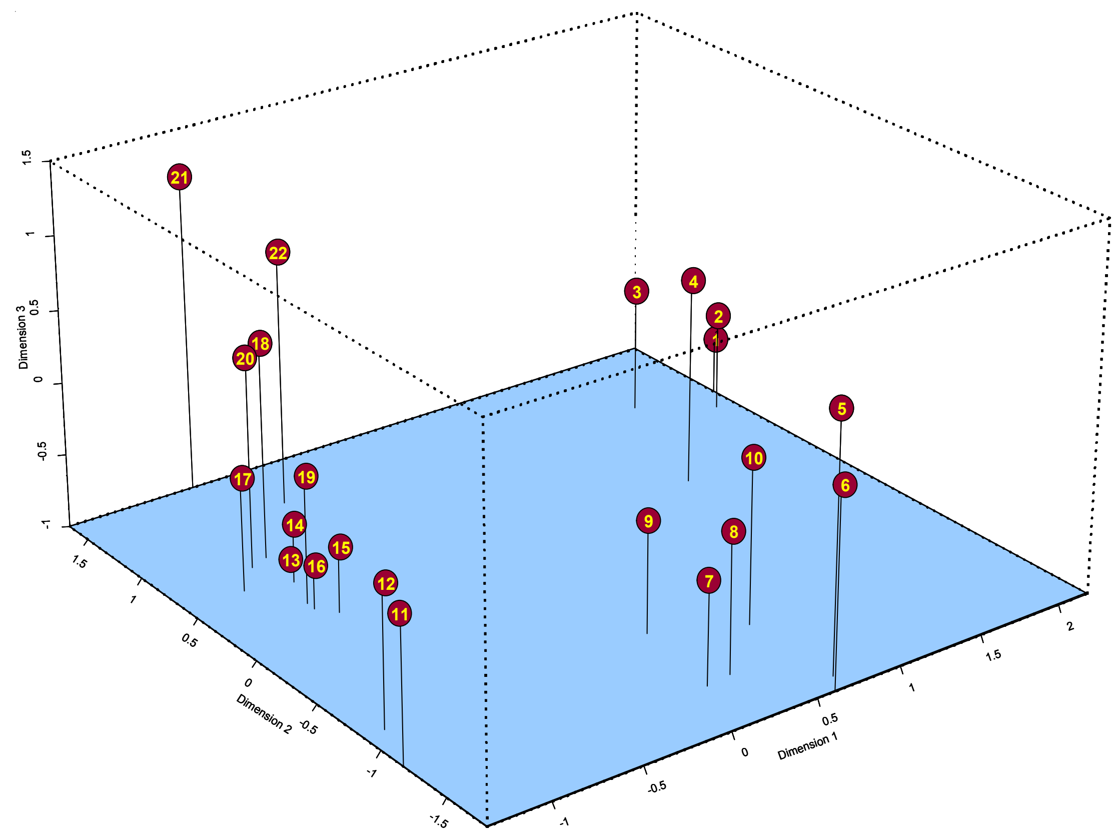
Sets of loci identified by the above methods were tested for usefulness for estimating relative contributions to mixed stock fisheries in the Yukon River. Using simulations in which the relative contribution of all stocks in the baseline were estimated for simulated mixtures composed entirely of fish from a single population (1000 iterations), we measured the accuracy and precision of population composition estimates based on the reduced sets of markers. The performance of the selected sets of markers were compared with the perfomance of sets of randomly selected markers.

Yukon River drainage and Chinook salmon populations



Population Structure

Using the delta distances calculated between populations and plotting these distances in three dimensions (multidimensional scaling analysis) we can display relationships among populations. The dots and numbers match the populations on the map and in the legend. Populations can be segregated into groups based on genetic and geographic factors.



Results: Locus Ranks

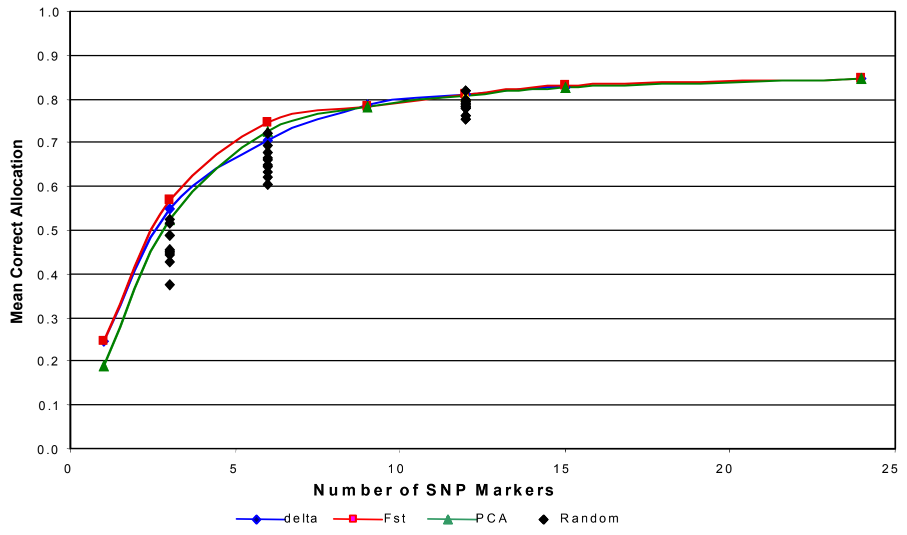
The three different methods of ranking the SNP markers did not sort the markers identically, but in general the "best" markers appeared near the top in each ranking. Twelve markers are common to all sets of the top fifteen ranked markers.

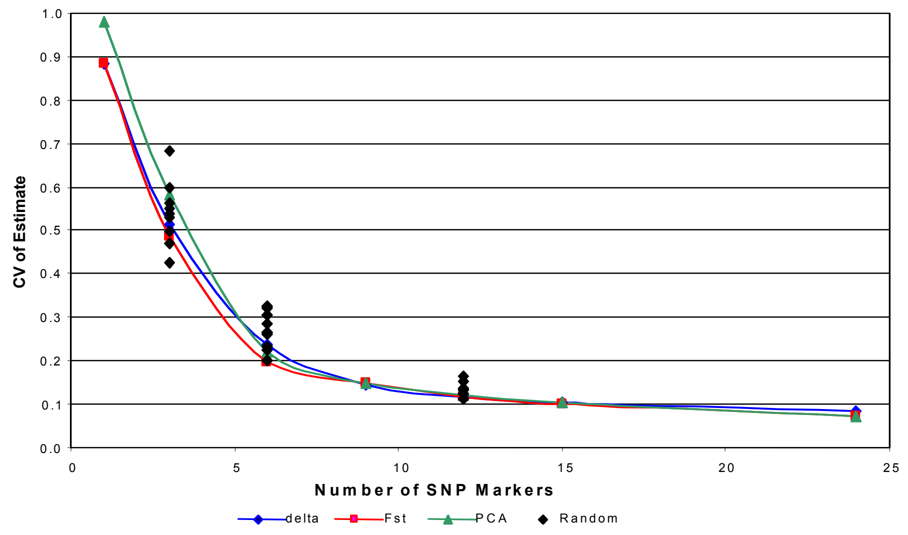
Spearman's rank correlation test indicated that the *delta* and the Fst ranks were more similar to each other (r=0.94; 90%CI [0.53-1.00]) than either was to the PCA ranking (r=0.87; 90%CI [0.46-1.00] and r=0.77; 90%CI [0.36-1.00], repectively).

SNP	Ranks		
Locus	delta	Fst	PCA
Ots_GH2	1	1	3
Ots_IGF-I.1-76	2	2	1
Ots_GPDH-338	3	4	4
Ots_P53	4	8	6
Ots_E2-275	5	5	8
Ots_Tnsf	6	3	7
Ots_SClkF2R2-135	7	6	9
Ots_Prl2	8	9	10
Ots_FGF6A	9	11	18
Ots_MHC1	10	12	13
Ots_u6-75	11	15	12
Ots_FGF6B	12	14	17
Ots_SWS1op-182	13	7	5
Ots_Zp3b-215	14	10	14
Ots_SL	15	16	11
Ots_P450	16	20	22
Ots_u202-161	17	13	2
Ots_Ikaros-250	18	18	20
Ots_Ots2	19	17	16
Ots_MHC2	20	19	15
Ots_LWSop-638	21	22	21
Ots_u4-92	22	21	19
Ots_ins-115	23	23	23
Ots_u211-85	24	24	24

Results: Stock Identification

The ranked sets of SNP markers show a rapid increase in mean population identification up to the top nine ranked markers, after which the addition of more markers demonstrate only small improvement (**Top Graph**). Perfect identification is 100% correctly assigned to the contributing group. All sets of "best" markers outperformed the randomly chosen sets. The coefficient of variation of the mean estimate dropped very rapidly from one to six marker sets (**Bottom Graph**). All three sets of ranked markers performed similarly and each was more accurate than randomly chosen sets ofmarkers. The variation in the estimates with sets of three "best" markers was the same as with random sets, but precision improved more rapidly for chosen sets than for random sets as more markers were added.





Conclusions

- The three methods of ranking the SNP markers by information content provided similar ranks.
- The ranked sets of SNP markers performed more accurately and with better precision than randomly chosen sets of markers for mixed stock analysis.
- Only a relatively small set of SNP markers (24) was available. This process may show greater differences between the ranking methods and improved mixed stock analysis performance when more SNP markers are available for analysis.

Acknowledgements

Funding for this study was provided by the Yukon River Salmon Research and Management Fund and the State of Alaska. Assistance in this analysis was provided by L. Seeb, C. Smith, N. DeCovich, A. Barclay, and Z. Grauvogel.