

# Evolution of a Perfect Simple Sequence Repeat Locus in the Context of Its Flanking Sequence

Scott M. Blankenship,\* Bernie May,† and Dennis Hedgecock\*

\*Bodega Marine Laboratory and †Department of Animal Science, University of California–Davis

Microsatellites, which have rapidly become the preferred markers in population genetics, reliably assign individual chinook salmon to the winter, fall, late-fall, or spring chinook runs in the Sacramento River in California's Central Valley (Banks et al. 2000. *Can. J. Fish. Aquat. Sci.* **57**:915–927). A substantial proportion of this discriminatory power comes from *Ots-2*, a simple CA repeat, which is expected to evolve rapidly under the stepwise mutation model. We have sequenced a 300-bp region around this locus and typed 668 microsatellite-flanking sequence haplotypes to explore further the basis of this microsatellite divergence. Three sites of nucleotide polymorphism in the *Ots-2* flanking sequence define five haplotypes that are shared by the Californian and Canadian populations. The *Ots-2* microsatellite alleles are nonrandomly distributed among these five haplotypes in a pattern of gametic disequilibrium that is also shared among populations. Divergence between the winter run and other Central Valley stocks appears to be caused by a combination of surprisingly static evolution at *Ots-2* within a context of more rapidly changing haplotype frequencies.

## Introduction

Microsatellites are tandem repeats of short nucleotide sequence motifs. They have been enthusiastically adopted in the past decade for linkage and population genetic studies because their high polymorphism (Litt and Luty 1989; Tautz 1989; Weber and May 1989) is believed to resolve population structure better than allozymes (Bowcock et al. 1994; Blouin et al. 1996; Estoup et al. 1996; Jarne and Lagoda 1996). Although these highly variable loci are certainly a boon for individual assignment, pedigree, or parentage analysis, as well as for mapping studies, their use in classical analysis of population genetic structure has shortcomings. Hedrick (1999) showed that estimates of population differentiation may be erroneously low because microsatellite loci have high within-population heterozygosity or may be inflated because of recent reductions in population size. Additionally, a statistically significant result regarding differences in allele frequency between populations may not reflect biological significance because statistical power is often high for hypervariable loci (Hedrick 1999). Queney et al. (2001) reported that population level inference based on microsatellite genetic variation is also affected by time and space. Their nine microsatellite loci vastly underestimated the divergence between the main groups of European rabbit (*Oryctolagus cuniculus*) populations in Iberia, although they correctly identified the more recent colonization of France by *O. cuniculus*.

Microsatellites are thought to mutate predominantly by slippage of DNA polymerase during replication, which generally results in gains or losses of single repeat units, depending on the DNA strand in which the slippage occurs (Levinson and Gutman 1987b; Schlötterer and Tautz 1992; Weber and Wong 1993; Primmer et al.

1996; Wierdl, Dominska, and Petes 1997). The mutation mechanism for microsatellites appears consistent with the theoretical stepwise mutation model (SMM; Kimura and Ohta 1978) in which mutations are additions or subtractions of repeat units in the case of microsatellites. Convergent or recurrent types of mutations, although a fundamental part of SMM, are not consistent with the standard infinite alleles model (IAM, Kimura and Crow 1964), which assumes that every mutation that occurs within a population creates a unique allele. A slippage mechanism of mutation, which may occur commonly only in microsatellites because of their molecular structure, clearly has implications for inferences based on population phenotypic diversity because alleles can return to previous allele sizes or states, retarding the separation of allele frequency profiles between populations.

If microsatellites evolve in a stepwise fashion, convergence of unrelated alleles to a common size, size homoplasy, should be common at these loci, yet homoplasy has rarely been observed within populations (Estoup et al. 1995; Garza and Freimer 1996; Grimaldi and Crouau-Roy 1997; Culver, Menotti-Raymond, and O'Brien 2001). Homoplasy should also obscure the actual genetic distance between populations (Goldstein et al. 1995). This creates a paradox: microsatellites are useful because they are polymorphic, yet their mechanism of mutation obscures population differentiation by increasing homoplasy. Nevertheless, microsatellites do produce information generally concordant with other marker types, which suggests that homoplasy does not obscure differentiation of allelic frequencies. For example, microsatellites reliably discriminate the five major stocks of chinook salmon (*Oncorhynchus tshawytscha*) in California's Central Valley (Banks et al. 2000) and corroborate information provided by allozymes (Bartley et al. 1992; G. Winans and D. Teal, personal communication), mtDNA (Nielsen et al. 1994), and MHC class-II B (Kim, Parker, and Hedrick 1999; Hedgecock et al. 2001).

A substantial proportion of the power to discriminate California's chinook populations comes from the *Ots-2* locus, a simple dinucleotide microsatellite [CA]<sub>6-27</sub>

Key words: evolution, microsatellite, simple sequence repeats, single nucleotide polymorphisms, linkage disequilibrium.

Address for correspondence and reprints: Scott M. Blankenship, Bodega Marine Laboratory, University of California–Davis, P.O. Box 1192, Occidental, California 95465. E-mail: szscottb@yahoo.com.

*Mol. Biol. Evol.* 19(11):1943–1951. 2002

© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

with confirmed disomic, Mendelian inheritance (Banks et al. 1999). Of the 10 microsatellite loci surveyed in that study, the *Ots-2* microsatellite exhibits the greatest allele-frequency divergence among chinook salmon populations in the Central Valley of California (Banks et al. 2000). The Sacramento River winter-run chinook, listed under state and federal laws as endangered (NMFS 1994), has a high frequency of the 7-repeat allele at the *Ots-2* microsatellite, making it genetically more distinct from the other Californian populations. We sought to resolve the paradox of how microsatellites reveal population structure by sequencing nonrepeat regions flanking the microsatellite repeat regions of the *Ots-2* locus, thereby placing the changes occurring at a putatively rapidly mutating microsatellite locus in the evolutionary context of its more slowly evolving flanking sequence. Point mutations are of the order of  $10^{-9}$  to  $10^{-10}$  events per locus per generation (Li and Graur 1991, pp. 69–73), several orders of magnitude less than the mutation rate estimates for microsatellites, which lie between  $10^{-2}$  and  $10^{-5}$  events per locus per generation (Levinson and Gutman 1987a; Henderson and Petes 1992; Weber and Wong 1993). Banks et al. (1999) estimated the mutation rate for the *Ots-2* microsatellite at  $10^{-4}$  events per locus per generation. The flanking sequence haplotype in which each microsatellite allele is embedded can be used to determine the historical relationship among microsatellite alleles, and specifically for the pure repeat *Ots-2* locus, required for constructing genealogies. In previous studies, homoplasmy has largely been characterized by differences in the composition of complex repeat arrays (Estoup et al. 1995; Garza and Freimer 1996; Grimaldi and Crouau-Roy 1997; Angers, Estoup, and Jarne 2000). In this study, if two microsatellite alleles of the same size differ at one or more flanking nucleotides, we assume that these alleles are homoplasious, having converged by stepwise mutation to a common repeat number along separate flanking sequence lineages.

For multiple chinook populations, the *Ots-2* tandem repeat array and 300 bp of associated flanking sequence were isolated for individual alleles. The sequence flanking the microsatellite repeat contains polymorphisms; hence, the sequence information for a chromosomal haplotype contains two loci: the tandem repeat and the flanking sequence with its complement of flanking single nucleotide polymorphisms (SNPs). In this study we compare the observed flanking sequence variation and the distribution of microsatellite repeats at the *Ots-2* locus.

## Materials and Methods

### Population Samples

Three populations from the Central Valley of California were studied: winter, fall, and spring runs. Individuals used in this study are a subset of those studied by Banks et al. (2000). Winter-run samples were from wild-caught individuals used as broodstock in the United States Fisheries and Wildlife Service Coleman National Fish Hatchery (CNFH). Individuals collected

from 1991 through 1995 were used. Additional winter-run samples were also obtained from CNFH in 1998. Fall-run samples were also obtained from individuals used as broodstock at the CNFH. Fall samples comprised individuals collected from 1993 and 1994. Genotype frequencies within hatchery samples conform to random mating expectations and the populations show no evidence of relatedness (Banks et al. 2000). Spring samples are from the California Department of Fish and Game collections of wild individuals from Deer Creek, during 1996–1998, and Mill Creek, during 1995 and 1998. An “outgroup” of chinook salmon from the Quesnel River is from a 1996 collection of wild individuals supplied by the Canadian Department of Fisheries and Oceans. DNA was extracted from all samples, as described by Banks et al. (1999, 2000).

### Primer Development and PCR

We developed a new primer based on the sequence of the genomic DNA clone from which *Ots-2* was originally developed (Banks et al. 1999). The new primer, 2A(2)-R, 5'-GTC AGG AGT AAC TTT AT-3', replaced the original primer *Ots-2-R* and was used in combination with the previously described PCR primer, *Ots-2-L*, 5'-ACA CCT CAC ACT TAG A-3' (Banks et al. 1999) to amplify a fragment encompassing as much known flanking sequence as possible (GenBank accession #AF107030). The program Oligo 4.0 was used to select the exact primer sequence (Applied Biosystems). The new PCR primer, 2A(2)-R, and the original primer, *Ots-2-L*, amplified a fragment containing the microsatellite and 300 bp of flanking sequence. The PCR components for this locus were 1.0 mM MgCl<sub>2</sub>, 0.05 mM dNTPs, 1.0 μM PCR primer, 0.025 U/μl *Taq* (Promega), and 50 ng template DNA. Temperature and cycling parameters were denaturing at 94°C for 30 s, annealing at 45°C for 15 s, and extension at 72°C for 20 s, repeated for 35 cycles on a tetrad thermocycler (MJ Research). Amplified fragments were separated by electrophoresis on 8% denaturing polyacrylamide gel, and fluorescently end-labeled amplicons were visualized using the FMBIO II imaging system (Hitachi Software Engineering America Ltd.).

### Evaluation of Flanking Haplotypes

Two methods were used to isolate haplotypes for sequencing purposes: subcloning and gel isolation. Because homozygous individuals were common, haplotypes were isolated predominantly by subcloning (95%), although some haplotypes were obtained through gel isolation (5%). Cloning of alleles was completed as described subsequently. PCR-amplified fragments were run onto a 2% agarose gel. Bands, visualized using ethidium bromide, were excised from the gel. DNA was recovered from the agarose and purified using QIA-Quick purification columns (Qiagen). Purified DNA was cloned using the TOPO TA\* cloning kit (Invitrogen). For each subclone, a minimum of six minipreps was performed. Plasmid DNA was isolated using either standard protocols (Sambrook, Fritsch, and Maniatis 1989,

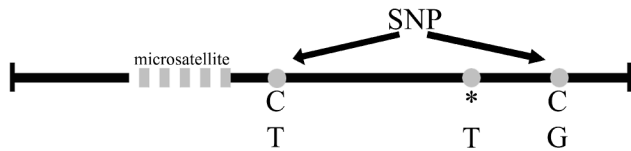


FIG. 1.—Schematic of a haplotype comprising a microsatellite tandem repeat and a collection of flanking polymorphisms, two SNPs and one insertion-deletion. Cytosine (C), Thymine (T), indel (\*), Guanine (G).

pp. 1.23–1.28) or the Qiaprep kit (Qiagen). We sequenced cloned DNA using either the fmol<sup>®</sup> cycle sequencing kit (Promega) and fluorescently end-labeled primers or templates were sent to Davis Sequencing (Davis, CA). Gel isolation of haplotypes was also accomplished for a small number of individuals. PCR-amplified fragments were separated on 8% denaturing polyacrylamide gel and stained with SYBR<sup>®</sup> Gold nucleic acid stain (Molecular Probes). Stained fragments were visualized using the fluorescent imager. By aligning reference lines drawn on the gel plate with the gel image, individual bands were located in the gel for excision. Isolated bands were placed in 20  $\mu$ l of ddH<sub>2</sub>O and incubated overnight at 4°C. Ten microliters of a 1:50 dilution of supernatant was used as template for 100  $\mu$ l PCR reactions. Reaction products were concentrated using QIAQuick purification columns (Qiagen) and sequenced as described previously.

### Genotyping

For the region flanking the microsatellite repeat array, sequence variation is referenced to the original clone sequence (GenBank accession #AF107030). A combination of polymorphisms is considered a flanking sequence haplotype. Genotype information for each allele at the *Ots-2* locus consists of a flanking haplotype and a microsatellite repeat array (fig. 1). Discrepancies in the length of the microsatellite repeat array occurred for some individuals because of sequence stutter. As a result, clones from a problematic sample could contain multiple sequences for repeat size (e.g., 8-, 9-, 10-repeat units). In such cases, cloned DNA was reamplified using PCR, and repeat length was independently determined through standard polyacrylamide electrophoresis. If a haplotype could not be reliably determined for an allele it was discarded; because both alleles were not identified in all individuals, sample sizes are not necessarily even numbers.

### Within-Population Analysis

The *Ots-2* microsatellite frequencies obtained from winter, fall, and spring population samples were compared, using chi-square tests and *F*-statistics, with published data from Banks et al. (2000). Published *Ots-2* allele frequency data are not available for the Quesnel River population. A chi-square test for homogeneity between frequencies observed in this study and previously published frequencies was performed using pseudoproportionality tests of significance (Zaykin and Pudovkin 1993). In addition, a pseudoproportionality test was used to com-

pare microsatellite allele frequency profiles for a haplotype across all populations. To compare observed allele frequency profiles and data from Banks et al. (2000), Wright's *F*-statistics were also calculated using the program Genetix version 4.0 (Belkhir et al. 2001). Tests for random mating proportions of genotypes within samples (H-W-C equilibrium) were calculated for each locus (repeat array; SNP haplotype) using the program Genepop, version 3.1 (Raymond and Rousset 1995). Linkage disequilibrium was measured using the program Genetix, version 4.0.

### Between-Population Analysis

To obtain estimates of genetic differentiation between populations, statistics based on different mutation models, Wright's *F*-statistics and Slatkin's  $R_{ST}$  were calculated. Wright's *F*-statistics, based on the IAM, were calculated using the program Genetix, version 4.0. Calculations were made using microsatellite allele frequencies and frequencies of flanking sequence haplotypes. Additionally, because each microsatellite has a known flanking sequence haplotype, a composite allele was generated combining information from both loci. The composite locus was used to evaluate homoplasy, the occurrence of the same sized microsatellite with different flanking haplotypes. *F*-statistics were calculated using the composite alleles, and the estimates derived from this no-homoplasy (at the resolution of the data) alternative were compared with the estimates calculated for the flanking and the microsatellite loci alone. The significance of estimates was assessed using 1,000 permutations in Genetix. A measure of population differentiation assuming an SMM  $R_{ST}$ , was calculated using the program RST CALC, version 2.2 (Goodman 1997). The estimate was calculated using only the microsatellite frequency data because the repeat array is the only locus that would be consistent with a stepwise model. All estimates were calculated using four population groupings, one using all four sample populations, one using only Central Valley stocks, one excluding winter-run, and one comparing fall and spring-run populations.

## Results

### Flanking Variation

Two SNPs and one insertion-deletion are observed among the 668 alleles sampled. Referenced to the original clone sequence (GenBank accession #AF107030), position 180 contains a C or T, position 323 contains a single base deletion (\*) or T, and position 383 contains a C or G (see fig. 1). Five haplotype combinations of the three segregating sites are observed: (1) CTC, (2) C\*C, (3) T\*C, (4) T\*G, and (5) TTC. Four haplotypes are possible, given the three polymorphic sites; hence, a rare back mutation or recombination event is required for the formation of a fifth haplotype. Although the flanking haplotypes are shared across all populations, the frequencies of the flanking sequence haplotypes differ among populations (table 1, parts a–d). Of the 199 winter-run fragments analyzed, 0.75 are haplotype-1, 0.17 are haplotype-2, 0.06 are haplotype-3, 0.01 are hap-

**Table 1**  
**Flanking Sequence Haplotype microsatellite Allele**  
**Frequency Matrices for (a) Winter Run (b), Fall Run (c)**  
**Spring Run and (d) Quesnel River**

TANDEM REPEATS	HAPLOTYPES					Total
	1	2	3	4	5	
<b>(a) Winter</b>						
7....	146	5				151
8....						—
9....	4	27	1			32
10....						—
11....			2			2
12....						—
13....						—
14....					1	1
15....						—
16....			1	1		2
17....		2	5	1		8
18....						—
19....						—
20....						—
21....						—
22....			2		1	3
23....						—
24....						—
25....						—
26....						—
27....						—
Total	150	34	11	2	2	199
<b>(b) Fall</b>						
7....	7					7
8....						—
9....	4	46	7			57
10....		1				1
11....	3	3	5			11
12....						—
13....			2			2
14....	1	2	14	1		18
15....			1			1
16....			7			7
17....		6	29	5		40
18....						—
19....						—
20....						—
21....						—
22....			8			8
23....						—
24....						—
25....			1		1	2
26....						—
27....			1			1
Total	15	58	75	6	1	155
<b>(c) Spring</b>						
7....	11	1				12
8....						—
9....		63	4			67
10....						—
11....	1	1	3			5
12....						—
13....						—
14....			10			10
15....						—
16....			8	1		9
17....		1	20	6		27
18....						—
19....						—
20....						—
21....						—
22....			7			7

**Table 1**  
**Continued**

TANDEM REPEATS	HAPLOTYPES					Total
	1	2	3	4	5	
23....						—
24....			4	1		5
25....			3			3
26....			1			1
27....						—
Total	12	66	60	8	0	146
<b>(d) Quesnel River</b>						
7....						—
8....						—
9....	1	96	1			98
10....						—
11....	2	7	39			48
12....						—
13....						—
14....			6	2		8
15....						—
16....			6			6
17....			8			8
18....						—
19....						—
20....						—
21....						—
22....						—
23....						—
24....						—
25....						—
26....						—
27....						—
Total	3	103	60	2	0	168

lotype-4, and 0.01 are haplotype-5 (table 1, part a). This frequency profile differed from the other populations in which haplotypes 2 and 3 are the most common. Of the 155 fall-run fragments analyzed, the frequencies of haplotypes 1 through 5 are 0.10, 0.37, 0.48, 0.04, and 0.01, respectively (table 1, part b). Of the 146 spring-run fragments analyzed, the frequencies of haplotypes 1 through 4 are 0.08, 0.45, 0.41, and 0.06, respectively (table 1, part c). Of the 168 Quesnel fragments analyzed, the frequencies of haplotypes 1 through 4 are 0.02, 0.61, 0.36, and 0.01, respectively (table 1, part d). Haplotype-2 is the most common haplotype overall.

**Population Genetic Analysis—Within Population**

The distributions of *Ots-2* microsatellite alleles and the associations of each repeat size with flanking haplotype are shown in table 1, parts a–d, with relative two-way classification frequencies shown in figure 2. For example, the 7-repeat allele observed in the winter-run population has a frequency of 0.76, and a majority of those 7-repeat alleles are associated with flanking haplotype-1 (fig. 2).

Microsatellite allelic diversity differs by haplotype, with haplotype-3 containing the largest number of microsatellite alleles, followed by haplotype-2 (table 2). Disregarding run and pooling over all populations, the total number of microsatellite alleles observed in haplotypes 1–5 are 4, 6, 12, 4, and 3, respectively. Haplo-

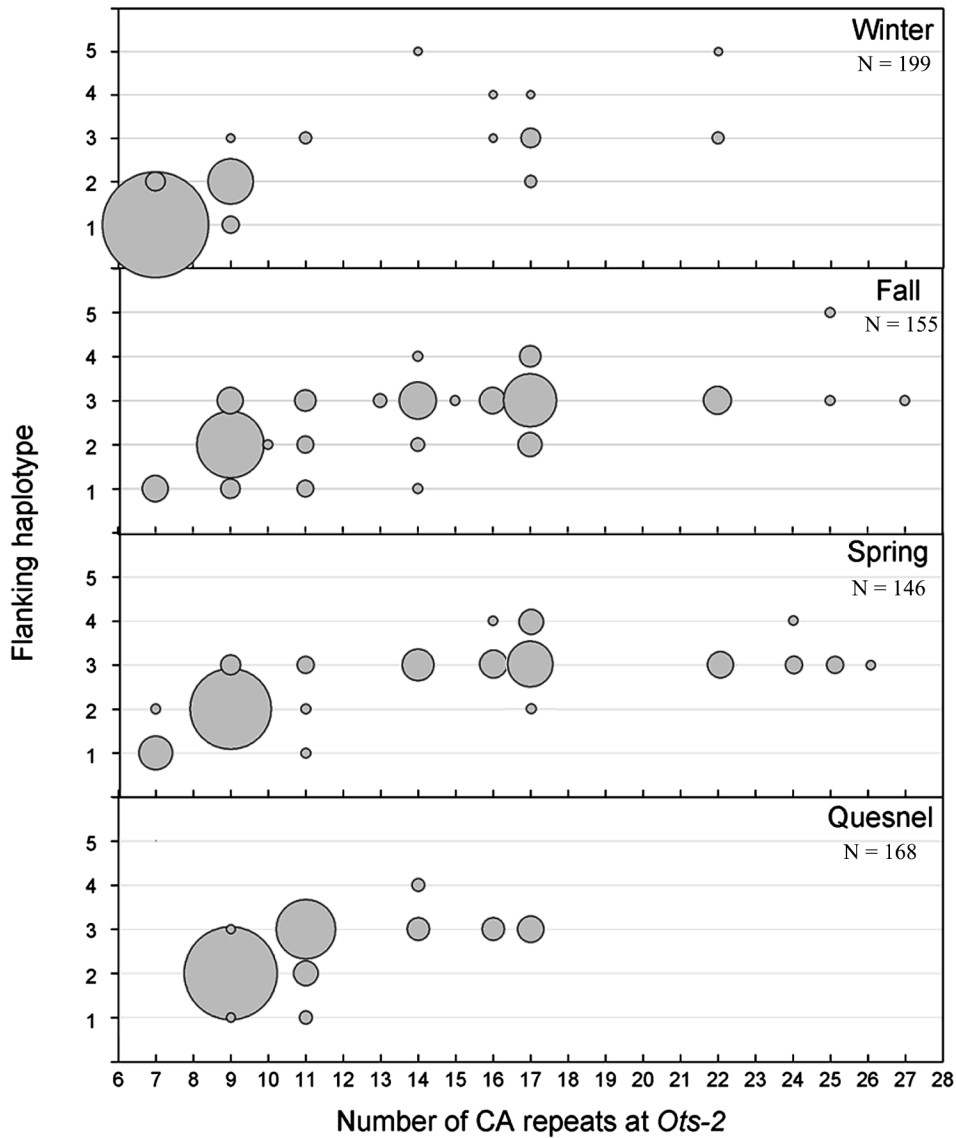


FIG. 2.—Frequency data of *Ots-2* microsatellite alleles categorized by haplotype. Position of gray bubble along x-axis denotes repeat size, position on y-axis signifies flanking haplotype associated with the given repeat, and size of gray bubbles corresponds to frequency of gene doses. Data were square-root transformed to make rare alleles visible.

type-3 has the highest combined estimate of allele size variance (16.423) and heterozygosity (0.809) at the microsatellite (table 2). Among populations, fall run has the largest number of microsatellite alleles, with haplotypes 1–5 having 4, 5, 10, 2, and 1 allele(s), respec-

tively (table 2). Allele size variance and heterozygosity are generally highest at the microsatellite in fall run (table 2).

Banks et al. (2000) has previously reported *Ots-2* microsatellite allele frequency distributions for the win-

**Table 2**  
**Allelic Diversity Categorized by Haplotype**

	HAPLOTYP 1			HAPLOTYP 2			HAPLOTYP 3			HAPLOTYP 4			HAPLOTYP 5		
	N	ASV	H	N	ASV	H	N	ASV	H	N	ASV	H	N	ASV	H
Winter. . . .	2	0.11	0.05	3	4.45	0.34	5	17.60	0.71	2	0.50	0.50	2	32.00	0.50
Fall . . . . .	4	4.60	0.67	5	6.56	0.36	10	14.33	0.78	2	10.57	0.28	1		
Spring. . . .	2	1.33	0.15	4	1.09	0.09	9	18.73	0.82	3	6.50	0.41	0		
Quesnel. . .	2	1.33	0.44	2	0.26	0.13	5	5.91	0.54	2	0.00	0.00	0		
Total . . . .	4	0.94	0.17	6	2.55	0.21	12	16.42	0.81	4	7.88	0.51	3	32.30	0.67

NOTE.—N is the number of microsatellite alleles observed, ADV is the allele size variance, and H is the heterozygosity. The column totals are estimates from data pooled over all populations.

**Table 3**  
**Summary of Within-Population Genetic Analysis**

POPULATION	N	$\chi^2$	$F_{ST}$	H-W-C		LD	
				$\mu$ Sat	Flank	$r$	Prob.
Winter . . .	199	0.94–0.97	0.00	0.24	0.43	0.26	0.00
Fall . . . . .	155	0.0–0.001	0.01	0.23	0.36	0.15	0.04
Spring . . .	146	0.03–0.05	0.00	0.09	0.27	0.24	0.00
Quesnel . .	168	—	—	0.36	0.53	0.35	0.00

NOTE.—N is the number of microsatellite flanking sequence haplotypes determined;  $\chi^2$ , 95% confidence interval for the probability of homogeneity between observed microsatellite frequencies and data from Banks et al. (2000);  $F_{ST}$ , Wright's  $F_{ST}$  statistics comparing microsatellite allele frequency data from observed populations with published data from Banks et al. (2000);  $\mu$ sat is the  $P$  value for random mating proportions at the microsatellite locus; flank is the  $P$  value for random mating proportions at the SNP haplotype;  $r$  is correlation coefficient from Genetix (Weir 1979); Prob. is probability of observed value.

ter, fall, and spring populations. Our analysis uses a subset of individuals from Banks et al. (2000), and the microsatellite allele frequencies obtained are statistically homogeneous with published results. Using a chi-square pseudoprobability test of significance, the 95% confidence intervals around probabilities of homogeneity between published and observed data are shown in table 3. The winter allele frequency profile is almost identical, the spring profile is marginally homogeneous by chi-square, and the fall allele frequency profile is not homogeneous by chi-square analysis. For the spring and fall data, the population subsample is sufficiently different from the larger data set published in Banks et al. (2000) that an additional comparison was made using  $F$ -statistics (table 3).  $F_{ST}$ -statistics shown in table 3 reveal that minimal population substructure exists when population samples from this study are compared with published population data from Banks et al. (2000). No allele frequencies are published for the *Ots-2* locus in the Quesnel River. The genotype frequencies at both loci are consistent with the H-W-C expectations (table 3). Estimates of linkage disequilibrium at the *Ots-2* locus show a strong association between microsatellite allele and flanking sequence haplotype (table 3). Size homoplasy is inferred for most microsatellite alleles, with the 7-, 9-, 11-, 14-, 16-, 17-, 22-, 24-, and 25-repeat alleles being observed with two or more flanking haplotypes within a run (table 1, parts a–d).

#### Population Genetic Analysis—Between Populations

Estimates of population divergence calculated using  $F_{ST}$  and  $R_{ST}$  are shown in table 5 for the four population groupings. We estimated genetic divergence with and without winter run because the winter run is the most divergent population in the Central Valley (Banks et al. 2000).  $F_{ST}$  and  $R_{ST}$  estimates within all groupings are similar to each other (table 5).  $F_{ST}$  estimates for the flanking locus are most similar to  $R_{ST}$  for all the populations and Central Valley comparisons (table 5). The  $F_{ST}$  estimates using the composite locus tend to be lower than the estimates for other loci. The  $F_{ST}$  estimate for the composite locus and all populations is 0.224, a value lower than estimates made for the separate loci (table 5).

**Table 4**  
**Pseudoprobability Tests Comparing the Microsatellite Allele Frequency Distributions, by Haplotype, Across Population**

	ALL POPULATIONS		CENTRAL VALLEY ONLY	
	$\chi^2$	Probability	$\chi^2$	Probability
Haplotype-1 . . .	93.26	0.00	58.78	0.00
Haplotype-2 . . .	55.80	0.00	27.05	0.00
Haplotype-3 . . .	108.21	0.00	19.10	0.59
Haplotype-4 . . .	16.63	0.05	5.89	0.46

A pseudoprobability chi-square test comparing the microsatellite allele distributions for each haplotype across populations presents a counterpoint to the observed between-population genetic differences at the *Ots-2* microsatellite (table 4). When all sample populations are included, the microsatellite allele frequency distributions by haplotype are not equivalent; however, when only the Central Valley populations are analyzed, haplotypes 3 and 4 have equivalent microsatellite frequency distributions (i.e., comparisons within haplotype are not significant in some cases). This contrasts markedly with the highly significant  $F_{ST} = 0.252$  for the *Ots-2* microsatellite locus by itself.

#### Discussion

We sought to resolve the paradox of how highly polymorphic microsatellite loci reveal population structure even though their mechanism of mutation should seemingly increase homoplasy and obscure allele frequency differentiation. Our study relies on the identification, in genetically divergent chinook salmon populations, of variation at a perfect, simple sequence microsatellite locus in the context of variation in the unique DNA sequence flanking the microsatellite. Studies of genetic diversity in chinook salmon consistently show Californian populations to be distinct from other North American salmon populations (Utter et al. 1989; Bartley et al. 1992; Ford 1998). These same studies suggest that Canadian populations are the result of colonization from diverse groups of Oregon and Alaskan salmon. The inclusion in this study of widely divergent populations enables greater inference from patterns of microsatellite variation than would be possible using Californian chinook in isolation. Additionally, the SNP haplotypes flanking the pure repeat microsatellite are required to define the historical relationships among the microsatellite alleles. The concordance among microsatellite allele-frequency profiles observed from distinct sample populations may reflect a historical or remnant phylogenetic relationship that has not been removed by mutations at the microsatellite locus.

Although a single marker measures population divergence imprecisely, data from the *Ots-2* microsatellite parallels information derived from all marker types. *Ots-2* microsatellite data show winter run to be a genetic outlier in California (Banks et al. 2000), and this result is corroborated by allozymes (Bartley et al. 1992; Winans and Teal, personal communication), mtDNA (Niel-

**Table 5**  
**Between-Population Totals. Divergence Among Chinook Salmon Populations at a Microsatellite ( $\mu$ sat), its Flanking Sequence, and the Composite Locus**

	$R_{ST}$ $\mu$ sat	$\mu$ sat	$F_{ST}$ Flank	Composite
All .....	0.28	0.25	0.28	0.22
Excluding Winter ...	0.12	0.06	0.03	0.05
Central Valley .....	0.33	0.25	0.31	0.22
Spring-Fall .....	0.00	0.01	0.01	0.01

sen et al. 1994), and MHC class II B (Kim, Parker, and Hedrick 1999; Hedgecock et al. 2001). In addition, microsatellite data from this study show Californian salmon populations to be more similar to each other than to the Quesnel population (Canada) (table 5). Although the *Ots-2* microsatellite is highly polymorphic, the variation is structured by flanking haplotype. The associations between haplotype and microsatellite allele are strikingly conserved across sample populations (fig. 2). Haplotype information from genetically distinct groups has led us to conclude that changes in microsatellite allele frequencies are caused by shifts in frequencies of contextual haplotypes and not by stepwise mutation in the repeat region. Recent work in human genetics also points to the importance of chromosomal haplotypes for resolving population structure (Rioux et al. 2001; Stephens et al. 2001).

That nonrandom structure exists in the data sets is not necessarily surprising; however, what is not expected is that data from all populations should differ from random association in the same way (fig. 2). For example, flanking haplotype-2 is the most frequent haplotype; yet, it most commonly occurs with the 9-repeat microsatellite, and flanking haplotype-1 is associated most commonly with repeat 7. The correlation of repeat size with haplotype class is also shown by the strong linkage disequilibrium observed in the data (table 3 and fig. 2). Additionally, if the total linkage disequilibrium estimate is partitioned into the contribution of each allele combination, there is an excess of large, rare microsatellite alleles in haplotype-3 and an excess of common, small microsatellite alleles in nonhaplotype-3 classes. This observation is associated with the greater diversity seen in haplotype-3, irrespective of haplotype frequency (table 2 and fig. 2). This pattern of variation is not expected under a rapid stepwise mutation process, which should rapidly generate microsatellite allelic diversity, filling in the gaps of the microsatellite profiles for each haplotype and homogenizing the microsatellite allele frequency profiles within each population (S. M. Blankenship et al., unpublished data). This observation suggests that the microsatellite mutation rate is much slower than that expected, slower than changes in the frequencies of flanking haplotypes to which the *Ots-2* microsatellite is linked.

Differences among populations are attributable primarily to differences in flanking haplotype frequencies. The correlation between microsatellite alleles and flanking haplotype, established by gametic disequilibrium,

could explain how microsatellite markers produce concordant information with other marker types because population differentiation exists at the haplotype level. Conversely, populations are not different or are much more similar within a haplotype. For example, the microsatellite allele-frequency profiles at haplotype-3 are statistically equivalent among Californian populations (table 4). It appears possible that demographic effects (e.g., genetic bottleneck) alter the frequency of the flanking sequence haplotypes, which subsequently alter the microsatellite allele-frequency distributions. Demographic changes that alter flanking haplotype profiles could also account for the reduction in microsatellite allelic diversity observed in the winter and Quesnel populations (table 1, parts a and d). We need to invoke a combination of slow mutation at the microsatellite locus and genetic drift or selection at the level of flanking sequence haplotypes to explain the pattern of variation seen at *Ots-2*.

That most microsatellite alleles have multiple flanking sequences confirms homoplasmy for size. Moreover, the pattern of homoplasmy is the same within and among sampled populations, with homoplasmy occurring at almost every allele. For example, allele 9 showed homoplasmy in every study population; however, it most commonly occurred in haplotype-2 (fig. 2). Other examples are 11-repeat and 17-repeat alleles, which show homoplasmy in 3 of 4 sample populations but most commonly occur in haplotype-3. This observation would not have been possible without the use of SNPs flanking the microsatellite repeat array. The older flanking sequence haplotypes can be used to define the evolutionary relationships among the microsatellite alleles, given that the unique noncoding sequence has a mutation rate of approximately  $1 \times 10^{-9}$  per locus/generation (Li and Graur 1991), several orders of magnitude slower than that of microsatellite repeat arrays. Homoplasmy, the recurrence of same-size alleles in different haplotypes, clearly does occur within populations for simple sequence repeat loci.

Population analysis is not affected by homoplasmy, however, because microsatellite allele-frequency profiles are not solely influenced by stepwise mutation. The slow mutation at the microsatellite creates linkage disequilibrium and counteracts homoplasmy by confining microsatellite alleles to haplotypes. Perhaps lower levels of homoplasmy present at the *Ots-2* microsatellite due to linkage result in greater diagnostic power for the locus. Data presented in this study show that  $R_{ST}$  and  $F_{ST}$  give similar estimates of population divergence (table 5), although the  $R_{ST}$  value calculated for the population grouping that excluded the winter-run population is substantially higher than the  $F_{ST}$  estimates. The within-population component of variance was large for the population grouping excluding winter run, which contributed to a high value for  $R_{ST}$  even though the between-population component of variance was low in this grouping. Estimates of  $F_{ST}$  may have been reduced by high within-population heterozygosity (Hedrick 1999). The composite locus, which had the highest observed heterozygosity, generally had lower genetic distance estimates.

Whether the pattern of variation seen at *Ots-2* is typical for microsatellite loci remains to be seen in future studies of other loci. For example, *Ots-2* is located on the pseudolinkage group Oii on the rainbow trout linkage map (Sakamoto et al. 2000), where recombination may be reduced. If *Ots-2* proves to be typical, then the usefulness of microsatellites for population genetics may come from associations with larger chromosomal fragments whose frequencies are responding more directly to demographic factors governing population differentiation. Microsatellites afford an easy entrée to genome variation and are clearly valuable for mapping and parentage studies; nevertheless, a shift toward SNPs in population genetics may provide a better insight into the evolution of populations below the species level.

### Acknowledgments

We thank the U.S. fish and Wildlife Service, California Department of Fish and Game, and the Canadian Department of Fisheries and Oceans for providing samples used in this study. We would like to thank Charles Langley for helpful discussions about the research and J. Beyer and N. Belfiore for helpful comments on the manuscript. We would also like to thank three reviewers for helpful comments on the manuscript. This study was supported by funds from the California Department of Water Resources.

### LITERATURE CITED

- ANGERS, B., A. ESTOUP, and P. JARNE. 2000. Microsatellite size homoplasy, SSCP, and population structure: a case study in the freshwater snail *Bulinus truncatus*. *Mol. Biol. Evol.* **17**:1926–1932.
- BANKS, M. A., M. S. BLOUIN, B. A. BALDWIN, V. K. RASHBROOK, H. A. FITZGERALD, S. M. BLANKENSHIP, and D. HEDGECOCK. 1999. Isolation and inheritance of novel microsatellites in chinook salmon (*Oncorhynchus tshawytscha*). *J. Hered.* **90**:281–288.
- BANKS, M. A., V. K. RASHBROOK, M. J. CALAVETTA, C. A. DEAN, and D. HEDGECOCK. 2000. Analysis of microsatellite DNA resolves genetic structure and diversity of chinook salmon (*Oncorhynchus tshawytscha*) in California's Central Valley. *Can. J. Fish. Aquat. Sci.* **57**:915–927.
- BARTLEY, D., G. A. E. GALL, B. BENTLEY, J. BRODZIAK, R. GOMULKIEWICZ, and M. MANGEL. 1992. Geographic variation in population genetic structure of chinook salmon from California and Oregon. *Fish. Bull.* **90**:77–100.
- BELKHIR, K., P. BORSA, L. CHIKHI, N. RAUFASTE, and F. BONHOMME. 2001. 1996–2001 GENETIX 4.02, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France.
- BLOUIN, M. S., M. PARSONS, V. LACAILLE, and S. LOTZ. 1996. Use of microsatellite loci to classify individuals by relatedness. *Mol. Ecol.* **5**:393–401.
- BOWCOCK, A. M., A. RUIZLINARES, J. TOMFOHRDE, E. MINCH, J. R. KIDD, and L. L. CAVALLISFORZA. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**:455–457.
- CULVER, M., M. A. MENOTTI-RAYMOND, and S. J. O'BRIEN. 2001. Patterns of size homoplasy at 10 microsatellite loci in pumas (*Puma concolor*). *Mol. Biol. Evol.* **18**:1151–1156.
- ESTOUP, A., M. SOLIGNAC, J. M. CORNUET, J. GOUDET, and A. SCHOLL. 1996. Genetic differentiation of continental and Island populations of *Bombus Terrestris* (Hymenoptera, Apidae) in Europe. *Mol. Ecol.* **5**:19–31.
- ESTOUP, A., C. TAILLIEZ, J. M. CORNUET, and M. SOLIGNAC. 1995. Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). *Mol. Biol. Evol.* **12**:1074–1084.
- FORD, M. J. 1998. Testing models of migration and isolation among populations of chinook salmon (*Oncorhynchus tshawytscha*). *Evolution* **52**:539–557.
- GARZA, J. C., and N. B. FREIMER. 1996. Homoplasy for size at microsatellite loci in humans and chimpanzees. *Genome Res.* **6**:211–217.
- GOLDSTEIN, D. B., A. R. LINARES, L. L. CAVALLIS-FORZA, and M. W. FELDMAN. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**:463–471.
- GOODMAN, S. J. 1997. R-ST Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Mol. Ecol.* **6**:881–885.
- GRIMALDI, M. C., and B. CROUAEU-ROY. 1997. Microsatellite allelic homoplasy due to variable flanking sequences. *J. Mol. Evol.* **44**:336–340.
- HEDGECOCK, D., M. A. BANKS, V. K. RASHBROOK, C. A. DEAN, and S. M. BLANKENSHIP. 2001. Applications of population genetics to conservation of chinook salmon diversity in the Central Valley. Pp. 45–70 in R. L. BROWN, ed. *Fish Bulletin 179: contributions to the biology of Central Valley Salmonids*. California Department of Fish and Game.
- HENDERSON, S. T., and T. D. PETES. 1992. Instability of simple sequence DNA in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **12**:2749–2757.
- JARNE, P., and P. J. L. LAGODA. 1996. Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.* **11**:424–429.
- KIM, T. J., K. M. PARKER, and P. W. HEDRICK. 1999. Major histocompatibility complex differentiation in Sacramento River chinook salmon. *Genetics* **151**:1115–1122.
- KIMURA, M., and J. F. CROW. 1964. The numbers of alleles that can be maintained in a finite population. *Genetics* **49**:725–738.
- KIMURA, M., and T. OHTA. 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. USA* **75**:2868–2872.
- LEVINSON, G., and G. A. GUTMAN. 1987a. High frequency of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res.* **15**:5323–5338.
- . 1987b. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**:203–221.
- LI, W.-H., and D. GRAUR. 1991. *Fundamentals of molecular evolution*. Sinauer, Sunderland, Mass.
- LITT, M., and J. A. LUTY. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**:397–340.
- NATIONAL MARINE FISHERIES SERVICE (NMFS). 1994. Endangered and threatened species; status of Sacramento River winter-run chinook salmon. *Fed. Reg.* **59**:440–450.
- NIELSEN, J. L., C. GAN, J. M. WRIGHT, and W. K. THOMAS. 1994. Phylogeographic patterns in California steelhead as determined by MtDNA and microsatellite analyses. *Calif. Coop. Oceanic Fish. Investig. Rep.* **35**:90–92.

- PRIMMER, C. R., H. ELLEGREN, N. SAINO, and A. P. MOLLER. 1996. Directional evolution in germline microsatellite mutations. *Nat. Genet.* **13**:391–393.
- QUENEY, G., N. FERRAND, S. WEISS, F. MOUGEL, and M. MONNEROT. 2001. Stationary distributions of microsatellite loci between divergent population groups of the European rabbit (*Oryctolagus cuniculus*). *Mol. Biol. Evol.* **18**:2169–2178.
- RAYMOND, M., and F. ROUSSET. 1995. Genepop (Version 1.2)—population genetics software for exact tests and Ecumenicism. *J. Hered.* **86**:248–249.
- RIoux, J. D., M. J. DALY, M. S. SILVERBERG et al. (31 co-authors) 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* **29**:223–228.
- SAKAMOTO, T., R. G. DANZMANN, K. GHARBI et al. (12 co-authors) 2000. A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. *Genetics* **155**:1331–1345.
- SAMBROOK, J., E. F. FRITSCH, and T. MANIATIS. 1989. *Molecular cloning a laboratory manual*. 2nd edition. Cold Spring Harbor Laboratory Press, New York.
- SCHLÖTTERER, C., and D. TAUTZ. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20**:211–215.
- STEPHENS, J. C., J. A. SCHNEIDER, D. A. TANGUAY et al. (28 co-authors) 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**:489–493.
- TAUTZ, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* **17**:6463–6471.
- UTTER, F., G. MILNER, G. STÅHL, and D. TEEL. 1989. Genetic population structure of chinook salmon (*Oncorhynchus tshawytscha*), in the Pacific Northwest. *Fish. Bull.* **87**:239–264.
- WEBER, J. L., and P. E. MAY. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**:388–396.
- WEBER, J. L., and C. WONG. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**:1123–1128.
- WIERDL, M., M. DOMINSKA, and T. D. PETES. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* **146**:769–779.
- ZAYKIN, D. V., and A. I. PUDOVKIN. 1993. 2 Programs to estimate significance of chi-square values using pseudo-probability tests. *J. Hered.* **84**:152–152.

BRIAN GOLDING, reviewing editor

Accepted July 15, 2002