1 Pacific Salmon: Identifying Genetically Distinct Populations

Salmon form an integral part of the economy and cultural heritage of northwestern North America. Historically, the Columbia River basin once supported an estimated 7 to 15 million fish that returned to spawn each year. Heavy commercial fishing started a decline in the population of Columbia River salmon in the 1870s, a trend that has been exacerbated by land use changes, dams along the river system, changes in the ocean environment, and continued harvest. The population of spring Chinook salmon in Columbia River is currently less than 3% of the population size compared to when Europeans first arrived in the area

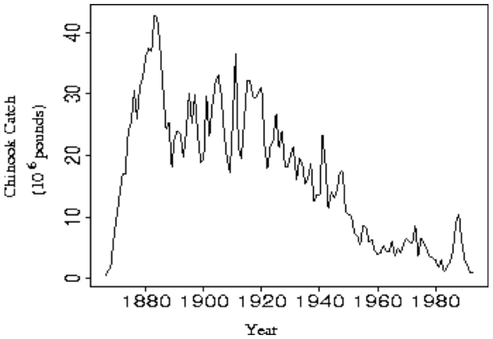


Figure 1. Population estimates of Chinook salmon in the Columbia River over the last century.

Certain populations of salmon in the Pacific Northwest have been listed as “threatened” or “endangered” under the US Endangered Species Act, so now their legal status governs much of the management decisions throughout the whole region. One of the key questions is to figure out which populations are protected and which are not.

The US Endangered Species act allows the listing of “distinct population segments” of vertebrate species. In the case of salmon the legal decision was to protect “Evolutionary Significant Units” or ESU. Some of those ESU are listed as threatened or endangered, while other populations are still thriving. To be considered an ESU, the population segment must be 1) reproductively isolated from other units and 2) contain a significant part of the evolutionary legacy of the species as a whole.

Migratory salmon are famous for their homing ability. After years in the ocean, adult salmon use faint chemical cues in the water to navigate upstream to spawn in the precise tributary where they were born. Thus there is very little mixing among populations. Although they migrate thousands of kilometers and mix freely in the ocean, when salmon return to spawn they will join other fish born in the same tributary and therefore will mate only with others from their particular population. There is a tiny bit of migration between streams. One estimate of the straying rate for hatchery fish was 1.2% but most of those move only 1 or 2 km from their site of birth. The result is that each drainage system will be genetically isolated from all others.

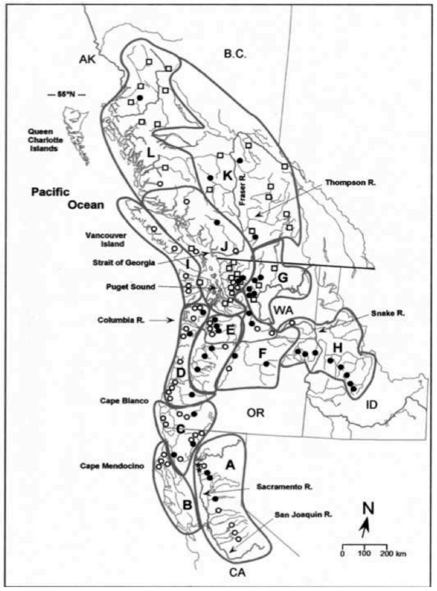
To the extent that adults return to their natal stream, each river or tributary can be considered an independent breeding population of salmon, with independent evolutionary history. Yet there is some mixing among populations in nearby streams.

How can we determine what the “evolutionary significant units” are?

Part of the answer will be genetic. By chance, reproductively isolated populations will accumulate genetic differences over time through mutation and random changes in allele frequency. Groups of populations that are genetically differentiated from others are likely to be the "evolutionary significant units".

1.1 Some Background (Allele frequencies and Hardy Weinberg Equilibrium):

Figure 2. Salmon in the Pacific Northwest can be divided into several genetically differentiated groups, or ESUs. associated with different drainage systems.



Mendel's rules describe the inheritance between individual parents and their offspring. Briefly, each individual receives two copies (alleles) of every gene, one from the mother and one from the father. Following meiosis, each gamete contains only a single copy of the gene. This is known as the **law of segregation** because the two alleles at a locus segregate into different gametes. The law of **independent assortment** says that inheritance of alleles at one locus is independent of what happens at other loci (as long as they are unlinked). Those rules of inheritance form the foundation of genetics.

What happens when we have a population of individuals (where Mendel's rules govern each individual mating)?

1.1.1 Allele frequencies

For most of what we do, we will assume that there is one locus with two alleles. Let's call the two alleles A and a, although the labels are arbitrary- we could also use A1 and A2 or any other naming convention. We can characterize the genetics at the population level by the frequency of each allele in the population, and the frequency of the genotypes in the population.

The traditional notation is to let p be the frequency of one allele and q be the frequency of the other. Because there are only two alleles those frequencies must add to 1.[1]

$$p+q=1$$

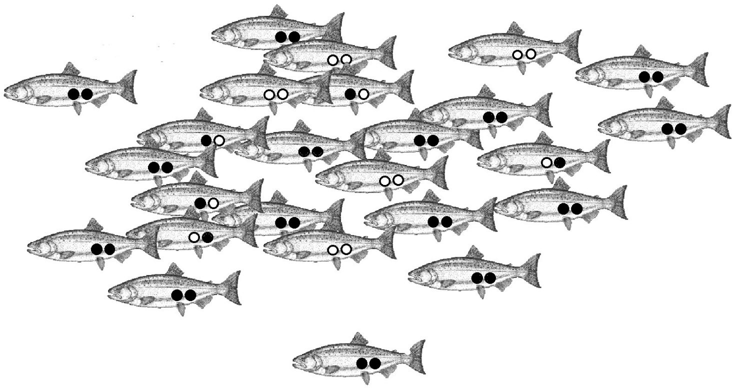
With two alleles, there are three possible diploid genotypes at this locus. If the two alleles are labeled A and a, then the three possible genotypes are: AA, Aa, and aa. We will use upper-case letters to specify genotype frequencies: P and Q for the two homozygotes and H for the heterozygotes. Again, because there are only 3 possible genotypes, the genotype frequencies must sum to 1:

$$P+H+Q=1$$

[1] Notice that population geneticists use the term "frequency" of alleles to mean the proportion, not the number, of alleles.

1.1.2 Consequences of Random Mating

One way to define random mating is that each gamete has an equal probability of combining with any other gamete to produce a zygote in the next generation. We will imagine a system of random mating where individual male and female gametes enter a common gamete pool and then are randomly combined to produce diploid genotypes for the next generation. This model is biologically accurate for some organisms like barnacles that have external fertilization. For barnacles, both male and female gametes are released into the water where they combine to produce new diploid offspring. However the model also works (mathematically) as a model of random mating for all organisms if the population is moderately large.



Let's imagine a population of 25 salmon. Each diploid parent has two copies of a pigment gene. The allele B produces black pigment, and the allele b is white. This particular population contains 15 BB homozygotes, 5 Bb heterozygotes, 5 bb homozygotes. Each fish is diploid, so there is a total of 50 copies of that gene in this population of 25 fish: 35 copies are of the black allele B and 15 copies are of the white allele b. Therefore the frequency of the B allele is 35/50 = 0.70.

Genotype	BB	Bb	bb
Number of fish	15	5	5

The frequency of an allele in a population is simply the proportion of all alleles that are of that type. Thus the easiest way to calculate the allele frequency is to use the method of "gene counting": simply sum the copies of that allele in the sample and divide by the total number of alleles. In general, if there are N individuals there will be 2N total alleles. Let p represent the frequency of the allele B. There will be two identical copies in each homozygote and 1 copy in each heterozygote. So the allele frequency can always be calculated as:

$$p = \frac{2*N_{BB} + N_{Bb}}{2N}$$

similarly,

$$q = \frac{2*N_{bb} + N_{Bb}}{2N}$$

In the example above, what is the frequency of the "b" allele?

q= ____________________

Now let's assume that these fish mate randomly to produce offspring in the next generation. (For now we won't keep track of males and females and just assume that the fish are all identical).

1.1.3 What offspring genotypes will be produced? How will the allele frequencies change?

Remember that gametes are haploid, so each gamete will have only one copy of the pigment gene. If the frequency of B in the population is 0.7, then 70% of the eggs will carry the B allele and 70% of the sperm will carry the B allele.

.

Under random mating, where the gametes can combine in all possible combinations, we can create a table like the one below to see how the eggs and sperm will combine. In this table, 70% of the rows are the B allele and 70% of the columns are the B allele. The offspring phenotype will be black if they inherit two copies of the B allele, white if they inherit two copies of b, and gray if they inherit one of each.

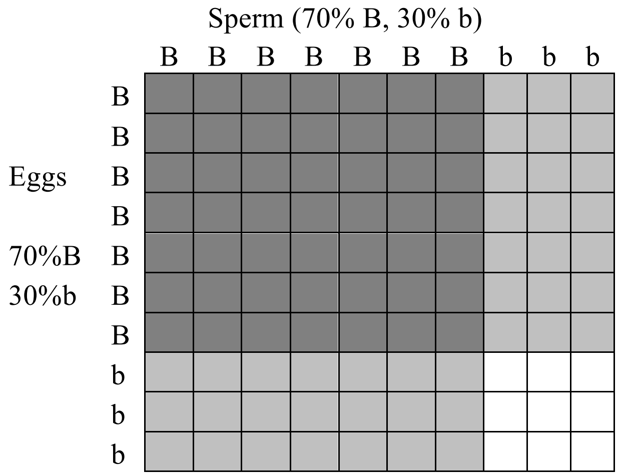


Figure 3.

After random mating the resulting offspring genotypes are: 49/100 BB, 42/100 Bb, 9/100 bb or in other words *P*=0.49, *H*=0.42, *Q*=0.09. (remember, the upper-case symbols are the genotype frequencies and lower-case symbols are allele frequencies).

Notice that the heterozygotes can be formed two different ways: Bb or bB.

These genotypes frequencies after random mating are a simple function of the allele frequencies. If p is the frequency of the B allele in the parents, then frequency of BB genotypes among the offspring is p^2: (0.7*0.7=0.49). The frequency of heterozygotes among the offspring is 2*p*q (2*0.7*0.3=0.42). And the frequency of bb genotypes is q^2 (0.3*0.3=0.09).

What are the new allele frequencies? If there are 100 squares representing offspring in the diagram above, there are 200 total alleles (because we are assuming that each individual is diploid, with one copy of each allele from their mother and one copy from their father). Each BB homozygote has 2 copies of B, whereas the heterozygote has only 1 copy.

total B alleles	n_B= 2*49 + 1*42 = 140
Frequency of B allele	p = 140/200 = 0.70
Total b alleles	n_b = 2*9 + 1*42 = 60
Frequency of b allele	q = 60/200 = 0.30

The final allele frequencies are 70% B and 30% b: exactly the same as the starting allele frequency. Random mating does not change allele frequencies at all.

The genotype frequencies, however, *did* change as the alleles combine to produce the new offspring. Our model of random mating will *always* produce genotype frequencies with the characteristic pattern of: p^2, $2pq$ and q^2 no matter what the genotype frequencies are prior to mating.

If a population starts with 80% B alleles (p=0.8) and 20% b alleles (q=0.2), what will be the allele frequencies and genotype frequencies after random mating?

New p = ________________

New q = ________________

Genotype frequencies will be: ______________________________

1.1.4 Assumptions for Hardy Weinberg Equilibrium (HWE)

The derivation above contains a profound result. In this very simple model that is based only the Mendelian laws of inheritance parents that mate randomly to produce the next generation offspring, *allele frequencies remain constant* from one generation to the next. And no matter what the starting genotype frequencies are, the offspring genotypes are produced in the *characteristic proportions* p^2*, 2pq and* q^2. This result is known as the **Hardy Weinberg Equilibrium**, named for the researchers who worked it out in the early 20th century.

Our explicit assumptions were simply that inheritance of alleles follows Mendel's laws and mating is random. No other evolutionary force was operating on the population.

Nevertheless, in that phrase "no other force" we are making four other implicit assumptions:

1. No mutation
2. No migration into or out of the population
3. The population is large (so the observed allele and genotype frequencies are exactly equal to the expected frequencies).
4. No fitness differences among individuals (all individuals have the same survival, and all matings produce the same number of offspring)

In later chapters, we will see how relaxing any of these assumptions can cause evolutionary changes. When none of those evolutionary forces are operating, we say that the population is in Hardy Weinberg Equilibrium.

Important: if the above assumptions are met, then genotype frequencies MUST be in HW proportions. If genotypes are not in HW proportions, then at least one of the above assumptions must have been violated.

That is a very strong result. HWE forms the basis of our understanding of the genetics of populations because it forms the null model against which all other evolutionary forces are compared.

Another consequence of HWE is that, at the population level, allele frequency is the starting point for characterizing the genetics of a species. Genotypes will change each generation as gametes from the two parents recombine in random combinations, but unless there is some external evolutionary force, the allele frequency of each population remains constant.

1.2 Differences in allele frequency define Evolutionary Significant Units

One way to define genetically distinct populations is to look for segments that have differentiated in allele frequency.

We have seen that the allele frequency is a fundamental unit for understanding the genetics of populations. All else being equal we expect the allele frequencies to remain constant in a given population unless some evolution force is acting. If two populations differ in allele frequency then they must have had unique evolutionary histories allowing the frequencies to diverge through some combination of drift or selection or mutation. Conversely, populations that are routinely linked by migration and mating are expected to have similar allele frequencies.

The following graph shows allele frequencies at just two loci for several populations of Chinook salmon in the Columbia and Snake Rivers. Spring run salmon are shown by open symbols and fall run salmon are dark symbols.

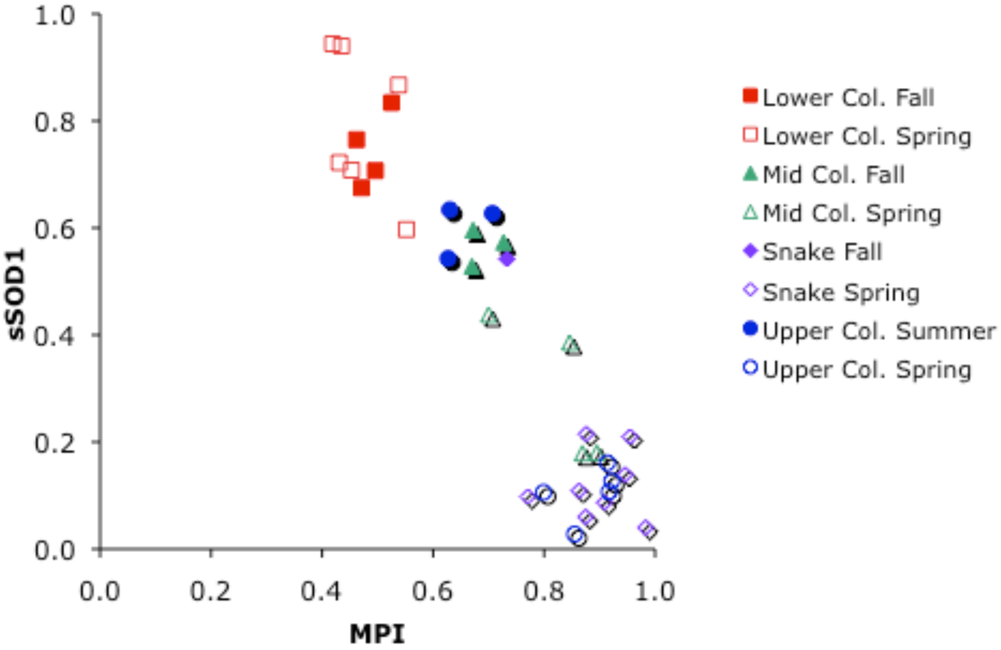


Figure 4. Allele frequencies for Mpi and SOD1 in eight samples of Chinook salmon from the Columbia and Snake rivers. These data show the frequency of the "fast" allele at each locus in each population. (data from Waples 2004)

From these data, what do you think are the distinct evolutionary significant units?

Should the spring- and fall-run Snake River salmon be treated as separate populations?

1.3 Population assignment: Where does this fish come from?

The various alleles can serve as genetic "tags" for different populations. Individuals from the tributaries will be more likely to have an allele if it is at high frequency in that particular population. Other fish from different populations will be likely to carry different alleles. The differences are not absolute: most alleles are present in all populations. But it is possible to calculate a *probability* that a fish comes from a certain population using the predictions from HWE.

That kind of population assignment procedure is currently used in the management of endangered salmon. For example, there is a commercial winter gill-net fishery in the lower Columbia river that targets primarily hatchery fish from the Willamette river. Those hatchery fish are available for harvest by the commercial fishermen. However, there are also an unknown number of salmon in the mouth of the Columbia that are from the threatened and endangered spring-run populations from the upper-Columbia and Snake Rivers. Those threatened populations are subject to strict harvest limits: legally only 4.1% of the upper river fish can be taken...The hatchery fish tend to arrive in the Columbia River mouth earlier than the upper-river fish. So the commercial season takes place from mid-February to March when the run is comprised mainly of lower Columbia River fish. But it is important to know when to stop the fishing in order to protect the upriver populations. When they approach the harvest limit for upper river fish, the season is closed to allow the upper river fish to pass through. How can we determine where a particular fish comes from ?

If we know the allele frequencies in the potential source populations, and if the source populations have genotype frequencies that are in HWE, then it is possible to calculate the probability that a particular genotype is produced by a particular source population.

Table 1. Average allele frequency for lower and upper river populations of Spring Chinook salmon, for a subset of 5 allozyme loci.

Locus	A	E	M	P	S
Lower Columbia & Willamette	A=0.829 a=0.171	E=0.831 e=0.169	M=0.471 m=0.529	P=0.796 p=0.204	S=0.740 s=0.260
Upper Columbia & Snake	A=0.997 a=0.003	E=0.080 e=0.920	M=0.876 m=0.124	P=0.157 p=0.843	S=0.821 s=0.179

This table shows the frequency of the "100" allele for each locus in the notation of Waples 2004. For simplicity we'll label this allele as an upper case A, E, M, P, or S depending on the locus. The frequency of the other (lower case) allele will be q=1-p.

Imagine we catch a fish, take a sample of tissue to the lab, and find that it has genotype AA at the A locus. What is the probability that populations from the lower Columbia will produce a fish with genotype AA? If the population is in HWE and p is the allele frequency for the A allele,, then p^2 of the fish should be homozygous AA. For the lower Columbia populations, the probability that randomly chosen fish has genotype AA is 0.829 * 0.829 = 0.687.

Now, what is the probability that Upper Columbia river populations produce a fish with genotype AA? The allele frequency for allele A in the upper river populations is p=0.997, so (again assuming HWE) the probability that a randomly chosen upper river fish has genotype AA is

Prob(upper river genotype is AA) = ____________________.

From these data, there is a fairly high likelihood of finding a fish with genotype AA in both populations. It is somewhat more likely to be produced in upper river populations than in lower river populations. We can calculate the odds of it being a fish from the upper Columbia as 0.996/0.687 = 1.45. All else being equal, that AA fish is about 45% more likely to come from the upper Columbia river than from the lower Columbia river.

Now lets add data from a second locus. Imagine that that same fish has genotype ee at the E locus. We can again calculate the probability of genotype ee being produced in the lower vs. upper river populations. The expected frequency of ee homozygotes is $(1-p)^2$. In the lower river populations p=0.831, so q=0.169 and the probability of finding genotype ee is 0.169*0.169=0.028. In the upper river populations p=0.080, so the probability of finding genotype ee is 0.920*0.920=0.846. Given the observation at the fish has genotype ee at the second locus, the fish is 30 times more likely to be from the upper river than from the lower Columbia river.

We can do better than that. What is the probability of finding a fish that is **both** AA **and** ee in the two populations (we will label it's multilocus genotype AA/ee)? The "and" rule for combining probabilities says that the probability of both events is the *product* of the probabilities. Therefore, Prob (genotype is AA and ee) = Prob(genotype is AA)*Prob(genotype is ee). For the lower river, 0.687*0.028 = 0.019..

What is the combined probability if it comes from an upper river population?

Prob (genotype is AA and ee) = _________________

Using the combined genotype data, the odds of it being from the upper river is 0.841/0.019 = 44. Using both loci we can say that it is 44 times more likely to have come from an upper river population. Although it is *possible* that the a fish with genotype AA/ee could have come from either population, it is much more likely to have been produced in the upper river.

Some basic probability rules:

The "AND" rule (product rule**):** The probability of two independent events both occur is the product of the probabilities. Prob (A and B) = Prob (A) * Prob (B).\
Example: if the frequency of allele r is 0.3, then the probability that two alleles are both r is 0.3 * 0.3 = 0.09.

The "OR" rule (sum rule**)**: The probability that either of two *mutually exclusive* events will occur is the sum of the probabilities. Prob (A or B) = Prob (A) + Prob (B).
Example: an individual will be "homozygous" if they have *either* genotype AA or genotype aa. Therefore the probability that an individual is homozygous is: Prob (AA or aa) = Prob(AA) + Prob (aa).
If the events are not mutually exclusive, you need to correct for the probability that both might occur. Prob (C or D) = Prob (C) + Prob(D) – Prob(C and D).

The "NOT" rule: The probability that event A does not occur is 1.0 – Prob(A)
Example: if the frequency of allele b in a population is 0.4, then the probability that a randomly chosen allele is *not* b is 1.0 - 0.4 = 0.6.

We can add data from as many loci as we would like. In general, the probability of a particular multi-locus genotype being produced in a particular population will be the product of the probabilities for each of the loci. If there are a total of *m* loci and L_i Is the likelihood of that genotype being produced at locus *i*, then the combined probability is:

$$L = L_1 \cdot L_2 \cdot L_3 \cdot ..L_m$$

Biologists often use this to assign an unknown fish to a particular source population. Then general procedure is to assume the fish comes from one of the possible source populations and calculate the probability of producing the observed genotype in that population. The procedure is repeated for each of the possible source populations and then the fish is assigned to the population that has the highest probability of producing that genotype. As a rule of thumb biologists often assume that there must be at least a 10x greater likelihood for a particular population in order to be confident of assigning the fish to that source.

1.4 Assumptions for population assignment tests:

- All potential source populations are included.
- Allele frequencies in the source population are known without error.
- The source populations are in HWE. That allows us to calculate the probabilities of observing particular genotype from the population allele frequencies.
- The loci are unlinked. That allows us to use the "and" rule and calculate the combined probability as the product of probabilities at individual loci.

1.5 Back to the Lower Columbia fishery: .

One of the management objectives is to monitor the timing of the runs to determine which populations of fish are migrating through at different times. In particular, fisheries managers want to know whether the run consists of mostly upper river fish (which are protected) or mostly lower river fish. Imagine that biologists captured a sample of 8 fish at the mouth of the Columbia river in early March, took a biopsy of their muscle tissue, and determined their genotype at 5 loci.

- Using the baseline data on allele frequencies in the upper- and lower-Columbia populations in table 1, assign the last three fish to their probable area of origin. Approximately what fraction of this run comes from the upper river?

Sample	**Genotype**	**Likelihood if from Lower Columbia**	**Likelihood if from Upper Columbia**	**Assign to**
1	AA/Ee/MM/PP/Ss	0.01046	0.00082	Lower
2	AA/EE/Mm/PP/SS	0.08205	0.00002	Lower
3	Aa/Ee/MM/PP/SS	0.000613	0.00001	Lower
4	AA/Ee/mm/pp/SS	0.00122	0.00108	(uncertain)
5	aa/EE/Mm/pp/SS	0.00023	6×10^{-9}	Lower
6	AA/ee/MM/Pp/Ss			
7	Aa/EE/Mm/pp/Ss			
8	Aa/Ee/MM/PP/Ss			

There is uncertainty about the origin of one of the fish, which had approximately equal probability of coming from either the upper or lower river populations. Of the remaining seven fish, six could be assigned to the lower river. That suggests that about 85 % of the fish in this particular sample come from the Lower Columbia River populations.

1.6 **Your turn: A fish story?**

(based on a true story, but with simplified genetic data)

In 2001, Canadian fisheries officers confiscated samples of salmon from a cannery that they suspected was illegally processing salmon from the Fraser River in Canada. Many populations of salmon in the Fraser River are protected from commercial fishing. The owners of the cannery claimed that the fish were in fact legally caught in southeast Alaska (where populations are strong), but they had no documentation to support their claim. The law enforcement agents confiscated some samples of fish, extracted some DNA from each sample, and determined the genotype at several loci. Here are representative genetic data from several possible source populations, and the individual genotypes of 2 of the fish. Which side do you believe?

Table 2. Allele frequencies in some possible source populations for four loci. Call the four loci A, B, C, and D. The table gives the frequency of one allele, which we will label with upper case A, B, C, D. The frequency of the other allele (lower case a, b, c, d) can be computed as q=1-p.

	Allele frequencies			
Source Region	OTS2-135 (A)	OTS9-105 (B)	OKe4-246 (C)	Omy325-96 (D)
Fraser River	0.35	0.52	0.04	0.23
Southeast Alaska	0.78	0.51	0.41	0.37
Yukon River	0.93	0.75	0.46	0.3
Columbia River	0.3	0.67	0.51	0.02

Genotype of confiscated fish samples: (Upper case letters correspond to the allele whose frequency is given in the table above. Lowercase letters indicate the alternative alleles).

Sample	Genotype
Fish 1	aa/bb/cc/Dd
Fish 2	Aa/BB/cc/dd

Answers:

p 4: $q= (5*2 + 5) / (2*25) = 0.30$

p 5: allele frequencies don't change from random mating, so $p=0.8$ and $q=0.2$. genotype frequencies will be $0.8*0.8 = 0.64$, $2*0.8*0.2 = 0.32$, and $0.2*0.2=0.04$.

p 7: For these two loci there appear to be at least three distinct clusters of samples: Lower Columbia spring and fall samples, the Mid&Upper Columbia/Snake fall-run fish, and the Mid&Upper Columbia/Snake spring run salmon.

Because the spring and fall Snake River populations have distinct allele frequencies they can be considered separate breeding populations and separate ESUs.

p 8 Prob = $.997*.997 = 0.994$

p 9 Prob (AA and ee) = 0.841

p 10 fish 6 0.00054 vs. 0.0502 → upper; fish 7 0.00544l vs. 0.000002 → lower; fish8 0.004332 vs. 0.000004 → lower

p 11 Fish1 Prob if Fraser = 0.03 Prob if Southeast Alaska = 0.001 ; Fish 2 Prob if Fraser = 0.067 Prob if Alaska = 0.012 . Both fish were more likely to be produced in the Fraser River.