

Population Substructure

Laurence D. Mueller

Department of Ecology and Evolutionary Biology

University of California, Irvine

Irvine, California, U.S.A. 92697

I. Origins of Population Substructure

Simplicity in scientific theories is usually seen as a virtue and population genetics is no exception. Most discussions of the genetics of populations starts with the simplest description of a population as a very large, single collection of randomly mating individuals. From this simple description genetic properties of populations may be deduced. For instance genes with multiple alleles are expected to obey the laws of Hardy-Weinberg and linkage equilibrium if they are not subject to natural selection and a sufficient number of generations of random mating has occurred. However, many real populations do not fit this simple model. Often we find populations have barriers that prevent the exchange of genes between them. These may often be physical barriers like mountains, oceans, or simply great distances. In these circumstances members of a species are found in many different subpopulations that are genetically different and isolated from each other. The collection of genetically differentiated subpopulations is referred to as population substructure.

Suppose a large population some time in the past sent out immigrants that created two new populations that were isolated from each other and from the parental population (figure 1(a)). Even if we assume these two new populations were initially genetically identical we expect that over long periods of time, perhaps dozens or even thousands of

generations, these populations will become genetically different from each other. These genetic differences may arise due to completely random processes like genetic drift or they may arise due to natural selection that acts differently in the two localities. More likely genetic differentiation may be due to both processes.

The particular history of a population may in fact be quite complicated giving rise to a hierarchy of events that affects the genetic characteristics of the population today. Thus, a single population may subdivide and give rise to two new isolated subpopulations that differentiate over time before these then subdivide and give rise to four subpopulations that persist today (figure 1b). The present day ecology may help identify this hierarchy. Thus, subpopulations 1-4 (figure 1b) may be fish in four small streams. However, subpopulations 1 and 2 are in streams that join a common river as are populations 3 and 4. Additionally these two rivers may ultimately join a single lake. There are clearly many other complicated hierarchies and subdivisions that can give rise to substructure in natural populations.

The present day populations may be completely isolated from each other or they may exchange migrants (figure 1b). The group of populations that communicate with each other through the exchange of migrants are called a metapopulation. Migration of individuals between populations may have effects on both the genetic variation and long term persistence of a population.

II. Genetic Consequences of Population Substructure

It is often difficult to identify the boundaries of subpopulations or even know if they exist. Consequently, population geneticists are often confronted with samples of individuals that may come from one subpopulation or may be from many subpopulations.

It turns out that even if all the subpopulations obey simple population genetic rules like Hardy-Weinberg and linkage equilibrium a pooled sample from many subpopulations will not. The nature of these effects depend on whether we are looking at one locus or multiple loci.

A. Single locus

Suppose we are interested in genetic variation at single locus with two alleles, called A and a . If there is population substructure as in figure 1a then the frequency of A in populations 1, 2, and 3 will be p_1 , p_2 , and p_3 respectively. The average of these three allele frequencies is \bar{p} . If each subpopulation is in Hardy-Weinberg equilibrium then the frequency of AA homozygotes in the three populations is p_1^2 , p_2^2 , and p_3^2 respectively. Let the average of these three values be \bar{P} . The naïve population geneticist may then take samples from all three populations, thinking they are a single population, and compare the observed frequency of homozygotes (\bar{P}) to the Hardy-Weinberg prediction, \bar{p}^2 . This comparison would always result in the observed frequency being greater than the predicted, that is $\bar{P} > \bar{p}^2$. This is called the Wahlund effect and is named after the Swedish geneticist who first described it in 1928, Sten Gösta William Wahlund.

We can in fact make a more quantitative statement about the difference between the observed frequency of homozygotes in the pooled sample *vs.* the Hardy-Weinberg expectation. Just as we used the allele frequencies in the individual subpopulations to estimate the mean allele frequency we can also use these values to estimate the variance in allele frequencies, which in this example is equal to $\frac{1}{3} \sum_{i=1}^3 (p_i - \bar{p})^2$. If we call the variance σ^2 , then the magnitude of the Wahlund effect is given by, $\bar{P} = \sigma^2 + \bar{p}^2$. This last

relationship will hold no matter how many subpopulations we have included in our pooled sample. It also suggests that the excess of homozygotes in our pooled sample will be proportional to the variation in allele frequencies. When there is no variation, $\sigma^2=0$, we will observe the Hardy-Weinberg expectation.

B. Two or more loci

Consider a second locus with two alleles, B and b . The frequency of the B allele in our three subpopulations (figure 1a) are r_1 , r_2 , and r_3 . It is usual to characterize the genetics of populations at multiple loci by examining gamete frequencies. For the two-locus genetic example considered here there are four possible gamete types, AB , Ab , aB , and ab . If we let their frequency in population 1, say, be x_{11} , x_{21} , x_{31} , and x_{41} respectively then this population is said to be in linkage equilibrium if $D=x_{11}x_{41}-x_{21}x_{31}=0$. D is called the coefficient of linkage disequilibrium. Even if all subpopulations are in linkage equilibrium, a pooled sample will generally not be. The magnitude of linkage disequilibrium in a pooled sample will be equal to the covariance in the frequencies of the A and B alleles over all subpopulations. Thus, if subpopulations with high frequencies of the A allele tend to either have very high frequencies of B or very low frequencies of B the pooled subpopulations will show substantial linkage disequilibrium.

If the subpopulations come back into contact and mate at random it will take many generations for linkage disequilibrium to vanish. The magnitude of linkage disequilibrium will be reduced by a factor of $1-r$ each generation, where r is the recombination fraction between the two loci. At best this means that linkage disequilibrium will be cut in half each generation if the two genes are unlinked. If there are more than two loci then in addition to the two-locus measures of linkage

disequilibrium there are higher order measures of associations between trios of loci, quadruples etc. These higher order measures of association will also eventually vanish with continued random mating although they may initially increase in magnitude unlike the two-locus disequilibrium values.

If recontact between the subpopulations does not result in random mating but only an exchange of limited migrants between their immediate neighbors, linkage disequilibrium between a pair of loci will vanish but at a slow rate. This rate will depend on the number of subpopulations and the rate of migration. As an example suppose the three populations in figure 1a receive 5% of their breeding population from their adjacent neighbors. Even if the *A* and *B* locus are unlinked the linkage disequilibrium of the pooled population will decrease by only about 5% per generation.

III. Wright's *F* Statistic

While we have summarized the Wahlund effect as the observation of an excess of homozygotes in a population of pooled subpopulations, it can also be stated as a deficiency of heterozygotes in the pooled population. Sewall Wright developed a statistic that makes use of this result. Using the parameters defined in IIA above, Wright's

fixation index is defined as, $F = \frac{2\bar{p}(1-\bar{p}) - \bar{P}}{2\bar{p}(1-\bar{p})}$. This parameter ranges in value from 0 to

1. When there are no differences in allele frequencies between the constituent subpopulations, $F=0$. Alternatively, when the subpopulations are fixed for alternative alleles, so that there are no heterozygotes in the subpopulations, F achieves its maximum value, 1. For genes that are not subject to natural selection several precise predictions about the expected magnitude of F may be made. In these cases genetic drift is the major evolutionary force causing the differentiation of populations. For instance, populations

with a structure like figure 1a, and no migration between populations or mutation at the studied loci will exhibit a steady increase in the magnitude of F until it eventually reaches

1. F increases at a rate that depends on the size of the subpopulations.

Evolutionary forces like mutation and migration may prevent F from reaching 1. This is because the individual subpopulations will not become fixed for any allele since the alternative allele will be continually reintroduced. In the case of migration relatively low levels of migration will reduce the final value of F to just moderate values. If sufficient time goes by the forces of drift and migration should equilibrate, producing an equilibrium or constant value of F equal to, $\frac{1}{4Nm+1}$, where N is the effective size of the population and m is the migration rate. For example if a population receives just two migrants per generation (e.g. $Nm=2$) F will equilibrate at 0.11.

IV. Migration Between Subpopulations

Migration can clearly have a substantial impact on the extent of population substructure. Typically it is very difficult to estimate migration rates for most species. Even if it is possible to document the movement of individuals from one location to another, these movements will have no genetic effect if those individuals don't mate and have offspring. However, it is quite easy to gather extensive genetic information on most natural populations with a number of different molecular based techniques. In 1981 Montgomery Slatkin devised a simple procedure for estimating rates of gene flow from genetic data.

Slatkin's technique requires an estimate of the frequency of private alleles. These are alleles that occur in only one of the many subpopulations examined. If gene flow between populations is very low we expect private alleles to have greater frequencies

than when gene flow is high. Gene flow may be expressed as the product of effective population size and migration rate, Nm . As described in section III Wright's fixation index -and thus the relative level of population substructure- will depend on the value of Nm . In table I we see very high values of Nm for marine mussels that indicate very little population substructure. This seems reasonable since these organisms distribute their immature larval forms into the ocean and the larvae may be carried great distances by ocean currents before they settle and become adults. On the other hand the study of *Plethodon cinereus* included samples from the Southern United States in Louisiana and as far north as Quebec, Canada. The ability of small terrestrial salamander to traverse these distances is clearly limited. Accordingly the estimates of gene flow are quite low.

V. Population Structure and Gene Trees

The ability to collect detailed genetic data directly from DNA sequences in natural populations has opened up new ways of studying population substructure. Consider a particular DNA sequence in a plant or animal mitochondria. This allows us to ignore the complications of recombination in the arguments that follow. Each copy of this particular sequence or haplotype must have originated from a single copy sometime in the past. We can in fact use the techniques traditionally used for phylogenetic inference to construct gene trees that show the likely history of particular haplotypes back in time.

In figure 2 we have shown a hypothetical gene tree. A single population starts out initially with four individuals, each with a different haplotype. Over time two of these haplotypes go extinct, a and d , while the other two, b and c persist. Additionally a barrier is set up that subdivides the population into two subpopulations. Samples of individuals from each of these subpopulations will confirm their genetic separation and their true

status since one subpopulation will consist entirely of the *b* haplotype, the other the *c* haplotype. In practice one must have some means of sampling putative subpopulations and then the gene tree is compared to the sampling units to see if there is congruence.

As an example the gene tree for the freshwater spotted sunfish, *Lepomis punctatus*, is shown in figure 3. There is a major split in the tree that corresponds perfectly to the samples that were taken from Western and Eastern localities.

Bibliography

- Bermingham, E. and Avise, J. C. 1986. Molecular zoogeography of freshwater fishes in the Southern United States. *Genetics* **113**:939-965.
- Christiansen, F. B. 1989. The effect of population subdivision on multiple loci without selection. In: *Mathematical Evolutionary Theory*, M. W. Feldman, editor. Princeton University Press, Princeton, N.J.
- Feldman, M. W. and Christiansen, F. B. 1975. The effect of population subdivision on two loci without selection. *Genetical Research Cambridge* **24**:151-162.
- Hartl, Daniel L. 2000. *A Primer of Population Genetics*. Third Edition. Sinauer, Sunderland, MA.
- Slatkin, M. 1985. Rare alleles as indicators of gene flow. *Evolution* **39**:53-65.

Suggested Cross References

Effective Population Size

Linkage Disequilibrium

Random Genetic Drift

Table I. Estimates of gene flow (Nm) per generation in several different animal species.

Species	Nm
Marine mussel (<i>Mytilus edulis</i>)	42.0
Fruit fly (<i>Drosophila willistoni</i>)	9.9
Mouse (<i>Peromyscus californicus</i>)	2.2
Fruit fly (<i>Drosophila pseudoobscura</i>)	1.0
Pocket gopher (<i>Thomomys bottae</i>)	0.86
Mouse (<i>Peromyscus polionotus</i>)	0.31
Salamander (<i>Plethodon cinereus</i>)	0.22

Figure Legends

Figure 1. The origin of population structure. (a) Initially samples from a large source population create three new subpopulations. Initially these subpopulations are genetically identical or at least quite similar. Over time these populations become genetically differentiated due to random genetic drift, natural selection, or both. (b) There can be a hierarchy of sampling events. In this diagram the source gives rise originally to two subpopulations. These become differentiated over time and then subdivide into a total of four populations that continue to differentiate. The present day populations may be completely isolated or may exchange some migrants as a metapopulation.

Figure 2. A hypothetical gene tree. Originally four different haplotypes exist in a single population, *a*, *b*, *c*, and *d*. Over time each of these can either leave 0 descendants, in which case the line ends, 1 descendant, symbolized by a single line or 2 descendants, indicated by a split with two new lines. At some time the single population is split by a barrier, shown as a gray bar, into two isolated subpopulations.

Figure 3. The gene tree for mitochondrial haplotypes of *Lepomis punctatus*. The haplotypes are identified by different numbers whereas the geographic samples are represented by different symbols. None of the locales where clones 1-8 (eastern samples) were found contain clones 9-17 (western samples). (After Bermingham and Avise, 1986)





