GNT2 01197

a0005

Population Substructure

LD Mueller, University of California, Irvine, CA, USA

Glossary

d0005 **Metapopulation** Multiple populations linked through migration.

d0010 **Migration** When individuals disperse from one location to another, either seasonally adventitiously.

AU4

d0015 **Private alleles** Alleles that are found in only one subpopulation of a species.

d0020 **Subpopulation** A collection of individuals within a species that is partially or completely isolated genetically from other populations often due to geographic barriers.

d0025 **Wahlund effect** A reduction in the frequency of heterozygotes, relative to the Hardy–Weinberg expectations, when a population is subdivided.

s0005 Origins of Population Substructure

p0005 Simplicity in scientific theories is usually seen as a virtue and population genetics is no exception. Most discussions of the genetics of populations start 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.

p0010 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 that 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.

p0015 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 again subdivide and give rise to four subpopulations that persist today (Figure 1(b)). The present-day ecology may help identify this hierarchy. Thus, subpopulations 1–4 (Figure 1(b)) 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 ultimate join a single lake. There are clearly many other complicated hierarchies and subdivisions that can give rise to substructure in natural populations.

p0020 The present-day populations may be completely isolated from each other or they may exchange migrants (Figure 1(b)). The groups 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.

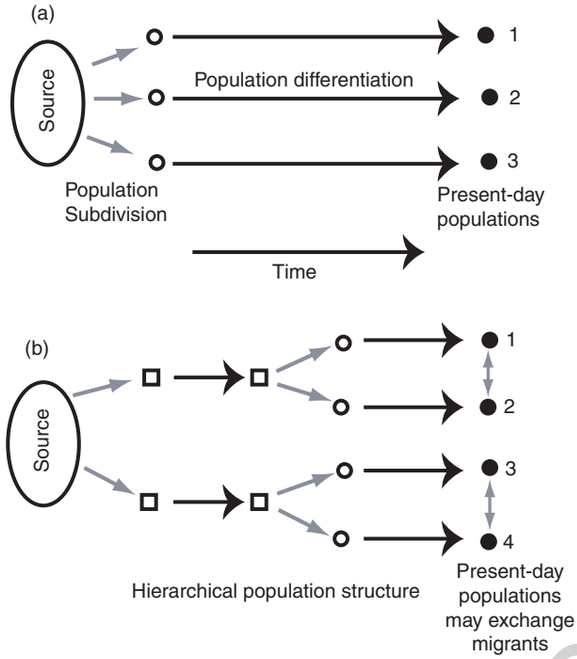
s0010 Genetic Consequences of Population Substructure

p0025 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.

s0015 Single Locus

p0030 Suppose we are interested in genetic variation at a single locus with two alleles, called A and a. If there is population substructure as in Figure 1(a), 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 naive population geneticist may then take samples from all three populations, thinking they are a single population, and compare the observed frequency of homozygotes ($\bar{P}$) with 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, Sten Gösta William Wahlund, who first described it in 1928.

p0035 We can in fact make a more quantitative statement about the difference between the observed frequency of homozygotes



f0005 **Figure 1** The origin of population structure. Black lines with arrows indicate the passage of time, whereas gray lines with arrows indicate movement of individuals. (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.

in the pooled sample versus 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 $\frac{1}{3}\Sigma_{i=1}^{3}(p_i-\bar{p})^2$. 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.

s0020 Two or More Loci

p0040 Consider a second locus with two alleles, *B* and *b*. The frequencies of the *B* allele in our three subpopulations (**Figure 1**(a)) 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.

p0045 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, and so on. 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.

p0050 AU2 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 slower rate. This rate will depend on the number of subpopulations and the rate of migration. As an example, suppose the three populations in **Figure 1**(a) 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.

Wright's *F* Statistic

s0025

p0055 Although 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 above, Wright's fixation index is defined as $F = (2\bar{p}(1-\bar{p}) - \bar{P})/(2\bar{p}(1-\bar{p}))$. 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 that shown in **Figure 1**(a), 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.

p0060 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

t0005 **Table 1** Estimates of gene flow (*Nm*) per generation in several different animal species

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

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 $1/(4Nm + 1)$, 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.

s0030 Migration between Subpopulations

p0065 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 do not 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.

p0070 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 previously, Wright's fixation index – and thus the relative level of population substructure – will depend on the value of Nm. In Table 1, 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 to 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 salamanders to traverse these distances is clearly limited. Accordingly, the estimates of gene flow are quite low.

Inferring Population Structure s0035

p0075 It has now become routine for population geneticists and evolutionary biologists to collect genetic data at many loci in a large number of individuals of a single species. Often, it is difficult to know how many subpopulations these individuals come from and if there is genetic exchange between these subpopulations. Jonathan Pritchard and his colleagues have devised a cleaver method for making these inferences from multilocus genetic data. Their method assumes that a sample of individuals comes from individual subpopulations that are themselves in Hardy–Weinberg and linkage equilibrium. The method, which relies extensively on computer simulations, then finds a population structure that is consistent with the assumptions of equilibrium in each constituent subpopulation. These methods rely on a computer-intensive statistical methodology called Markov chain Monte Carlo methods. In addition to providing an estimate of the number of subpopulations, these methods can also suggest the subpopulation membership of each individual in the sample as well as the fraction of an individual's genes that come from each subpopulation. AU3

Further Reading

Christiansen FB (1989) The effect of population subdivision on multiple loci without selection. In: Feldman MW (ed.) *Mathematical Evolutionary Theory*, pp. 71–85. Princeton, NJ: Princeton University Press. bib0005

Feldman MW and Christiansen FB (1975) The effect of population subdivision on two loci without selection. *Genetical Research, Cambridge* 24: 151–162. bib0010

Hartl DL (2000) *A Primer of Population Genetics*, 3rd edn. Sunderland, MA: Sinauer Associates. bib0015

Pritchard JK, Stephens M, and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959. bib0020

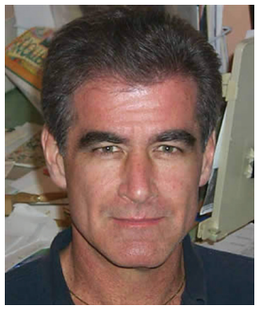
Slatkin M (1985) Rare alleles as indicators of gene flow. *Evolution* 39: 53–65. bib0025

Relevant Websites

http://pritch.bsd.uchicago.edu – Pritchard Lab, The University of Chicago – Structure. bib0030 AU5

GNT2 01197

Biographical Sketch



Laurence D. Mueller received his PhD from the University of California, Davis, and he did postdoctoral research at Stanford University before starting his first faculty position at Washington State University. He has been at the University of California, Irvine, since 1988 were he currently is professor of the Department of Ecology and Evolutionary Biology. His research interests are in life-history evolution, aging, and the population genetic aspects of forensic DNA typing. Dr. Mueller is the author of over 100 research papers in these fields as well as two books: *Stability in Model Populations* and *Evolution and Ecology of the Organism*.