N	μ	s	m	r
∞	-	-	-	-

Population Subdivision

So far, our considerations have been limited to a single population with random mating. However, real populations are spatially structured introducing a component of non-random mating: geographically close individuals are simply more likely to mate with each other. Geographic structure is often of interest in itself, as it can give away information on the populations' history (e.g. human colonization history). It can help conservation policy makers in defining **evolutionary significant units** (populations with an independent evolutionary history), guiding them in their decisions where conservation efforts should be directed. Moreover, population structure is a disturbance that needs to be taken into account when we aim to infer selection or demographic processes. Accurate description of population subdivision is thus a central theme in population genetics and constitutes the basis of most population genetic studies.

We start by describing population structure in the most simple case, with two a priori defined subpopulations 1 and 2 each in Hardy-Weinberg-Equilibrium that are isolated from one another and not connected by gene flow. Assume a diploid locus with alleles *A* and *a* with allele frequencies *p1* and *p2* for the *A* allele in subpopulation 1 and 2 respectively. The average allele frequency $\bar{p}$ across both subpopulations is given by the weighted average

$$\bar{p} = \frac{2N_1p_1 + 2N_2p_2}{2N_1 + 2N_1}$$

For equal population size this reduces to

$$\bar{p} = \frac{p_1 + p_2}{2}$$

For the latter case the average expected proportion of heterozygote individuals in the subdivided populations H_S is

$$H_s = \frac{2p_1(1-p_1) + 2p_2(1-p_2)}{2} = p_1(1-p_1) + p_2(1-p_2)$$

H_S corresponds to the mean heterozygosity averaging over the expected heterozygosity of each subpopulation. Disregarding population structure by assuming a single panmictic population with allele frequency $\bar{p}$= (p_1+p_2) / 2 the expected total heterozygosity H_T (of the pooled population) would be

$$H_T = 2\frac{p_1 + p_2}{2}\left(1 - \frac{p_1 + p_2}{2}\right)$$

Now let the difference between the allele frequencies in the two populations be $\delta = |p1 - p2|$. Adding this into the equation above (adding and substracting $\delta^2/2$), we find

$$H_T = p_1(1-p_1) + p_2(1-p_2) + \frac{\delta^2}{2}$$
$$H_T = H_S + \delta^2/2$$

If $\frac{\delta^2}{2} = 0$ then $H_T = H_S$, if the allele frequencies differ $H_T > H_S$, i.e. the sub- populations contain fewer heterozygous individuals that expected given by the pooled allele frequency (**Wahlund effect**). This result is general including more than 2 populations, populations of different size or multi-allelic loci.

We can now make use of the relationship above to quantify population subdivision based on the differences in allele frequencies between populations. Wright defined the population measure of

$$F_{ST} = \frac{H_T - H_S}{H_T} = 1 - \frac{H_S}{H_T}$$

F_{ST} thus conceptualizes genetic variation in terms of deviation from HWE as the relative decrease in heterozygosity comparing the expected heterozygosity for a combined population (H_T) to that expected within populations (H_S) assuming Hardy-Weinberg proportions. Remember from above that $H_S = H_T - \delta^2/2$. We can thus rewrite F_{ST} as a function of the 'Wahlhund effect' relating it directly to allele frequency differences between populations as

$$F_{ST} = = 1 - \frac{H_S}{H_T} = \frac{\delta^2/2}{H_T}$$

Note that its value does not only depend on the amount of difference in allele frequency between populations ($\delta = |p1 - p2|$), but also on the overall amount of genetic diversity in the populations H_T. The denominator standardizes the measure by the maximal limiting variance $H_T = 2\bar{p}(1-\bar{p})$. It is thus a relative measure of population differentiation reaching from 0 (for $H_T = H_S$) to 1 for $H_T > 0$ and $H_S = 0$, occurring if alleles are fixed between populations $p_1 = 1$, $p_2 = 0$ or $p_1 = 0$, $p_2 = 1$). This is an important insight in the context of genome scans, where genomic regions might differ in the degree of genetic differentiation simply due to differences in the mutation rate ($4N_e u$).

Population differentiation and genetic drift

We can now conceptualize F_{ST} within the framework of the Wright-Fischer model. In independently evolving populations of finite size descending from a single origin without intermigration it can be shown that F_{ST} increases with time by

$$F_{ST} = 1 - e^{-t/2N_e}$$

where N_e is the effective population size of each population and t is the time of divergence in generations. Remember from the previous lecture that using the approximation $e(x) \approx 1 + x$ for $|x| \approx 0$ with $x = 1/2N$ we obtained $e^{-t/2N_e} = \frac{H_t}{H_o}$. Under these assumptions note that H_S/H_T provides an estimate of H_t/H_0. F_{ST} thus increases proportional to the loss of heterozygosity due to genetic drift. For populations derived from one common population F_{ST} will range from 0 in early generations and approaches unity as allele frequencies diverge over

time through genetic drift. When also considering migration (see below) we are faced with a challenge. We do not know whether low F_{ST} values are due to high migration rates or simply short divergence time between populations that have not reached migration-drift equilibrium yet (if they ever will). F_{ST} alone does not allow us to distinguish between different models of population history.

F_{ST} is a member of a class of similar statistics developed by Sewall Wright that explain deviations from HWE. These so called **F-statistics** formalize the idea of inbreeding with respect to different levels. Considering only one population, we can similarly compare the heterozygosity of the subpopulation expected under HWE $2p_S(1\text{-}p_S) = H_S$ to the observed frequency of the *Aa* genotype $f_{Aa}=H_I$. This is known as the **inbreeding coefficient**

$$F_{IS} = \frac{2p_Sq_S - f_{Aa}}{2p_Sq_S} = \frac{H_S - H_I}{H_S} = 1 - \frac{H_I}{H_S}$$

The level of comparison is clearly important. While someone with German ancestry may not show any sign of inbreeding when compared to the German sub-population (F_{IS}), (s)he probably will with respect to the entire European population differing in allele frequency and hence in the expected heterozygosity (F_{IT}).

$$F_{IT} = \frac{2p_Tq_T - f_{Aa}}{2p_Tq_T} = \frac{H_T - H_I}{H_T} = 1 - \frac{H_I}{H_T}$$

Hence, the reduction in heterozygosity within individuals compared to that expected in the total population can be decomposed to the reduction in heterozygosity of individuals compared to the subpopulation, and the reduction in heterozygosity from the total population to that in the subpopulation (**Figure 1**).

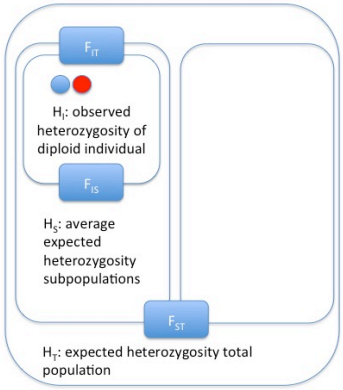

Figure 1: A diagram illustrating hierarchical F-Statistics. We can compare the average, obserced heterozygosity of individuals (H_I) to that found by randomly drawing diploid genotypes from the respective subpopulations (and their average H_S), to that found in the total population (H_T). Note that H_I is based on observed heterzygosity, H_S and H_T both refer to expected heterozygosity under HWE.

Subpopulation	Genotype frequency			Allele frequency		Expected H	F statistic
	AA	**Aa**	**aa**	**A**	**a**		
EXAMPLE 1							
Pop1	0.25	0.5	0.25	0.5	0.5	0.5	$F_{IS\_1} = 0.0$
Pop2	0.35	0.3	0.35	0.5	0.5	0.5	$F_{IS\_2} = 0.4$
	F_{IS}=0.2	F_{IT}=0.2	F_{ST}=0				
EXAMPLE 2							
Pop1	0.25	0.5	0.25	0.5	0.5	0.5	$F_{IS\_1} = 0.0$
Pop2	0.49	0.42	0.09	0.7	0.3	0.42	$F_{IS\_2} = 0.0$
	F_{IS}=0.0	F_{IT}=0.04	F_{ST}=0.04				

First example: we obtain H_S as the average of the expected heterozygosity of each population as $H_S = p_1(1-p_1) + p_2(1-p_2) = 0.25 + 0.25 = 0.5$. We then obtain H_T as the expected heterozygosity of the combined population as $2\bar{p}(1-\bar{p})$: $H_T = 2 \cdot 0.5 \cdot 0.5 = 0.5$. It follows that $F_{ST} = 1 - \frac{0.5}{0.5} = 0$ indicating no population differentiation. This is obvious, since pop1 and pop2 have the same allele frequencies. However, we can also calculate local inbreeding coefficients F_{IS_1} and F_{IS_2} for population 1 and 2 as follows: $F_{IS\_1} = 1 - \frac{0.5}{0.5} = 0$ and $F_{IS\_2} = 1 - \frac{0.3}{0.5} = 0.4$. While pop1 shows no deviation from HWE, the positive value for F_{IS} in pop2 indicates a strong deficiency in heterozygotes.

Second example: H_S is calculated as $H_S = 0.25 + 0.21 = 0.46$. We obtain $H_T = 2 \cdot 0.4 \cdot 0.6 = 0.48$. Accordingly $F_{ST} = 1 - \frac{0.46}{0.48} = 0.041\bar{6}$, $F_{IS} = 1 - \frac{(0.5+0.42)/2}{0.46} = 0$, $F_{IT} = 1 - \frac{(0.5+0.42)/2}{0.48} = 0.041\bar{6}$. Due to variation in allele frequencies $F_{ST}>0$. There is no evidence for global inbreeding at the subpopulation level (F_{IS} =0); yet, individuals are inbred with respect to the pooled population $F_{IT}>0$. Given the same magnitude for F_{IT} and F_{ST} we can conclude that this deviation from random mating is induced by population structure.

N	μ	s	m	r
finite	✓	-	✓	-

Finite population sizes and migration

While certainly useful, the above considerations are rather far from reality. In nature, population sizes are finite and populations generally exchange **migrants**, i.e. they are connected by **gene flow**. So what we are really interested in is to estimate the degree of differentiation under **mutation-migration-drift** equilibrium: mutation introduces variation to any of the populations, genetic drift will increase differentiation as allele frequencies are eventually driven to fixation, migration will homogenize allele frequencies and reduce differentiation. It turns out that in equilibrium the degree of differentiation as measured by F_{ST} is expected to be

$$F_{ST} = \frac{1}{1 + 4N_e(m+u)}$$

where *Ne* is the effective population size *u* is the mutation rate and *m* is the **migration rate** - the probability that an individual in the population is replaced by an incoming individual from

another population. Accounting for the fact that mutation rate tends to be orders of magnitude lower than the migration rate this reduces to

$$F_{ST} = \frac{1}{1 + 4N_e m}$$

However, for highly mutable markers, such as microsatellites, and in a situation of low gene flow this simplification should be given second thought.

The above result was derived by Wright using a particular population model, the **continent-island model**, assuming one large population sending out migrants to many (an infinite number actually of) small, unconnected satellite populations (often called sub-populations or **demes**). F_{ST} then measures the net effect of genetic drift in the finite population to constant migration rates under equilibrium. Despite of this artificiality this result is instructive, as it demonstrates that even a single effective migrant (N_em) is sufficient to reduce F_{ST} to almost zero. Assuming an effective population size of 10^5 as in humans very small migration rates indeed 10^{-5} are sufficient to homogenize allele frequencies among populations.

It is tempting to use this relationship to estimate the number of effective migrants N_em in natural populations by simply measuring F_{ST} from genetic marker data. However, we have to be cautious not to misinterpret N_em as a measure of gene flow. It is an artificial composite parameter, to infer gene flow we really would like to know the proportion of migrants m - which we generally don't unless we have access to an independent estimate of N_e. Also, in reality model assumptions are often not met, populations are not in equilibrium, or other forces such as selection increase allelic variance between populations.

The island model may be useful in a situation where there is a clear source (continent) population and receiving island populations. However, often populations are spatially assorted exchanging migrants more readily among adjacent populations. The **stepping-stone-model** accommodates for this reality and models migration rate as a function of geographic distance. Intuitively we would then expect low F_{ST} values for neighbouring populations and higher F_{ST} values for distant population comparisons. In fact, it turns out that

$$\frac{F_{ST}}{1 - F_{ST}} \sim geographic\ distance$$

This pattern is called **isolation by distance** and is commonly observed in natural populations.

What is a population?

In many cases we only have a vague idea of what constitutes a population in nature. Population identity in inferred from morphological characteristics or by broad-scale geographic features such as mountain ranges or rivers. Obviously, this can be highly misleading. In the following we will briefly consider the principle of two approaches that group individuals into genetic clusters by some similarity measure without having to predefine population structure.

Principle component analysis (PCA)

PCA is a multivariate method that summarizes information from many (often correlated)

sources into several synthetic variables, the principle components (PC). PCA has the desirable property that they decompose all the variance in the data into statistically independent (orthogonal) PC-axis explaining an ever-decreasing amount of the total variation. The use of PCA in population genetics was pioneered by Cavalli-Sforza in the 1970ies. With the increasing availability of genome-wide data sets with many millions of loci it has seen a significant revival during the last years. Originally proposed for summarizing genetic variance in allele frequency of predefined populations it has been expanded to individual genotypes. Importantly, PCA not only visualizes population structure in a graphically appealing way. Projections of principle components contain information on the genealogical history of a sample, and the genotype covariance matrix used for PCA is closely related to the allelic variance used in F-statistics.

Assignment methods

Originally, assignment tests were used to decide to which population a sampled individual belongs. The idea behind it is simple. Given a set of populations, and knowledge on the allele frequencies of those populations we can calculate the likelihood for each population for the genotype of the sampled individual would arise. E.g. population 1 has allele frequencies f_A=1 and f_a=0, population 2 has frequencies f_A=0.5 and f_a=0.5. Now consider an individual with genotype *Aa*. The probability Pr (*Aa*) is 0 coming from population 1 is 2x1x0=0 and from population 2 is 2x0.5x0.5=0.125. Hence we would conclude that the individual belongs to population 2. With multiple loci and differences in allele frequencies between populations, individuals can be assigned with high confidence to a population; even hybrids can be assigned to their most likely parental populations. This type of assignment still requires knowledge on population structure and the respective allele frequencies a priori. An important class of assignment methods has been developed to tackle this problem. These approaches allow the data themselves to determine if the population has subpopulations and, if so, how many would best explain the data at hand. The commonly used program *Structure*, for instance, uses Hardy-Weinberg equilibrium (HWE) assumptions to generate such hypothetical populations. Individuals of a sample are first randomly assigned to any of K clusters (populations). The algorithm then sequentially assesses deviations from HWE and shuffles the individuals such as to approach Hardy-Weinberg proportions. Maximum likelihood can be used to assess which number of clusters best represent the data. This is a very powerful approach allowing to 1) identify population structure and 2) assign the probability of population ancestry for each individual (**ancestry coefficients**).

Literature: (Barton et al. 2007; Futuyma 2013; Nielsen and Slatkin 2013)

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH. 2007. Evolution. 1st edition. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press

Futuyma DJ. 2013. Evolution. 3rd ed. Sinauer Associates

Nielsen R, Slatkin M. 2013. An Introduction to Population Genetics: Theory and Applications. Sunderland, Mass: Macmillan Education