Shane's Simple Guide to F-statistics

Intro

The aim here is simple & very focussed:

Aim : a very brief introduction to the most common statistical methods of analysis of population genetic structure (i.e., F-statistics and AMOVA), and how to interpret them.

Disclaimer!
Please remember that the following discussion is very simplified, describes only **one** of many approaches to F-stats, and ignores many assumptions and interrelated material. To ease understanding, some explanations are not 100% accurate, but should not be misleading. You will need to check text or original sources (or ask me) for further details and qualifications.

Background material

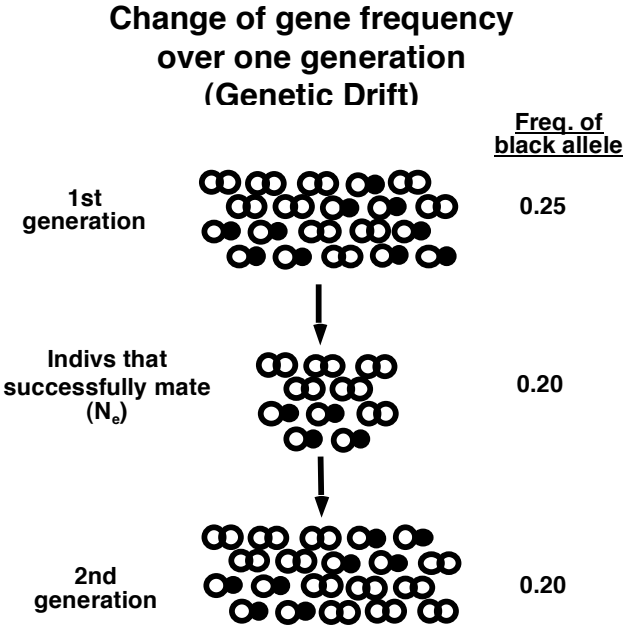
Terms

You need to familiar with what is meant by the following terms:

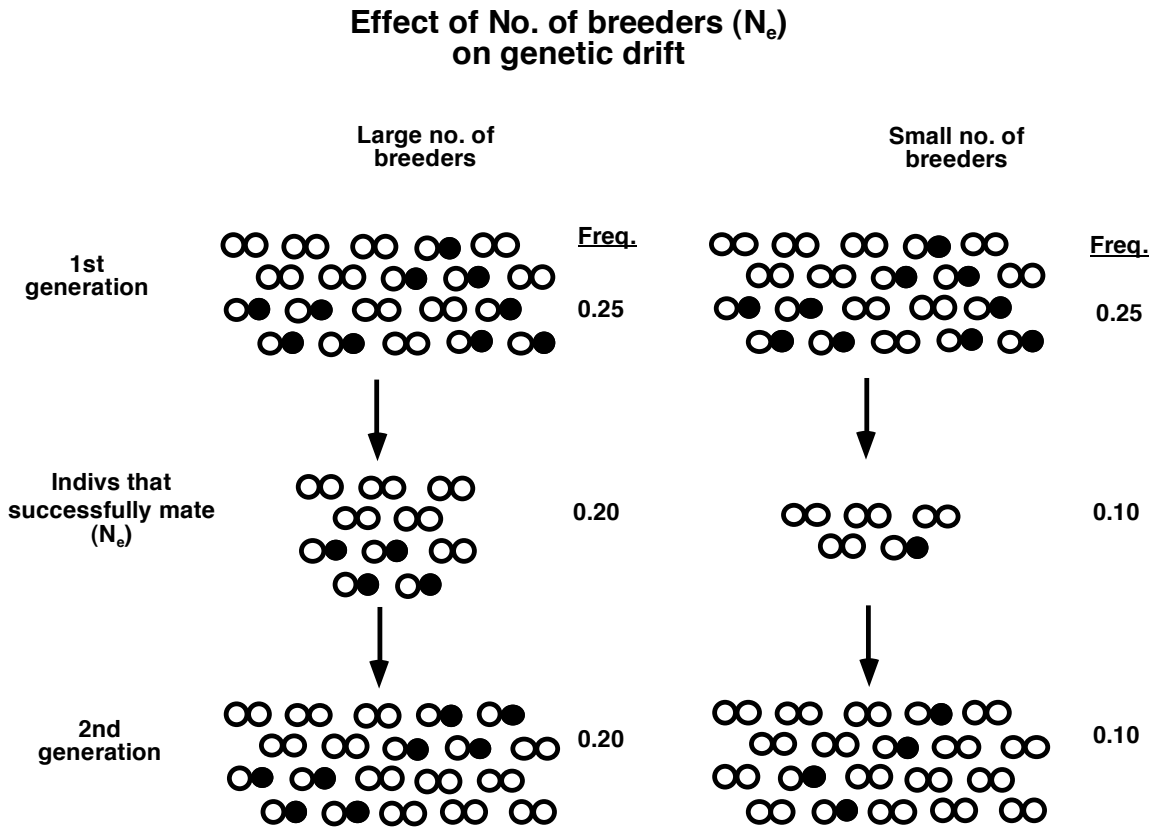
- heterozygosity
- homozygosity
- inbreeding – what it means in terms of genotype frequencies
- effective population size - N_e

Random genetic drift

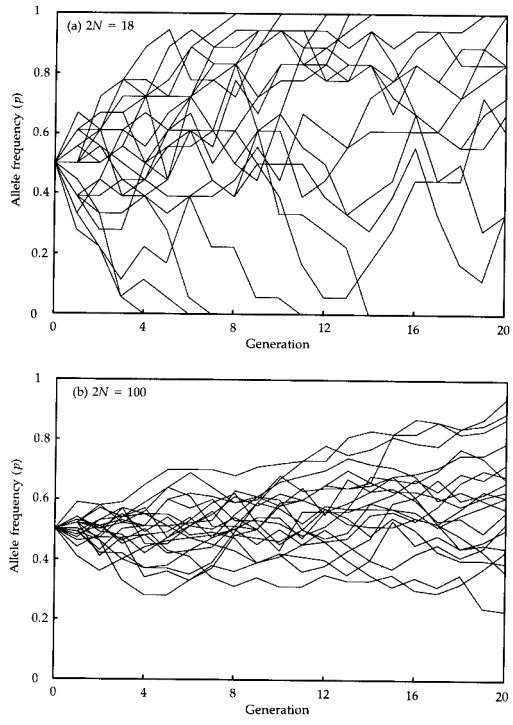
This is a central concept to understand in the genetic divergence of populations. To refresh your memory, one simple way of thinking of how genetic drift occurs from one generation to the next is shown below. (Note – where generations do not overlap, N_e can be thought of as the effective no. of breeders.) When only a small number of breeders contribute to the next generation, the small number of randomly-selected genes that are passed on are likely to differ slightly in their frequencies from the previous generation, simply due to the random sampling process.



The degree of genetic drift from one generation to the next depends on how large are the number of breeders (N_e). The smaller N_e , the larger the drift in frequencies from one generation to the next is likely to be:



The cumulative effect of this genetic drift over many generations is shown in the figure below. Each line shows the drift in frequency over time of one allele in one subpopulation. It can be seen that (with no migration) the populations gradually diverge genetically more and more over time. This effect is greater, obviously, for populations with smaller N_e.



Hardy-Weinberg Equilibrium

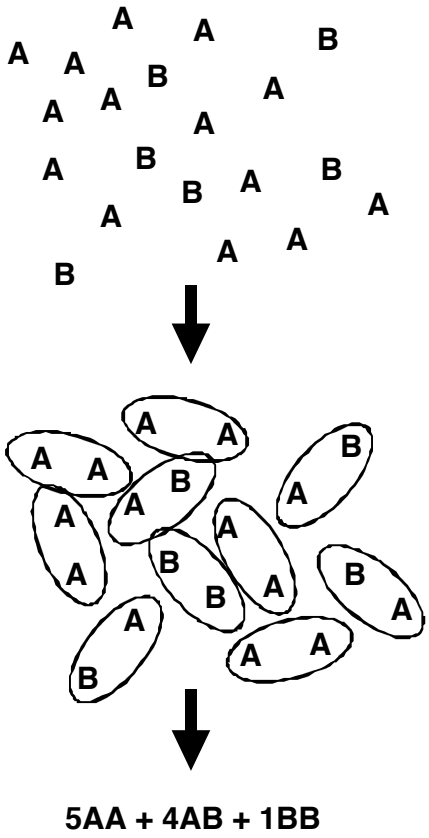
This is also a central concept in the derivation of F-statistics. The most basic point here (as illustrated below) is that, regardless of the genotype frequencies you start with in one generation, if there is completely random mating, then the genotype frequencies of the next generation will tend towards a highly predicted ratio - the HWE – that is determined entirely by the allele frequencies.

7AA + 3BB -> 14A & 6B

OR

5AA + 4AB + 1BB -> 14A & 6B

freq. [A] = p = 0.7
freq. [B] = q = 0.3



But how is this genotype ratio derived? Very simply, it comes from the probabilities of getting each of the three types of allele pairs (genotypes) shown above. Each of these three combined probabilities comes merely from the product of the probabilities (or frequencies) of the two alleles. This is shown below in both example numbers and in terms of the generalised allele frequencies, p and q. You might recognise that the formulae for the genotype frequencies are in the form of a binomial expansion. Thus, with more than two alleles the H-W frequencies can be easily determined using the appropriate expansion.

It can also be seen below that the formulae for calculating **expected** (as opposed to **observed**) heterozygosity and homozygosity come straight from the HWE. It is useful to note here that when there are more than 2 alleles, it is easier to calculate heterozygosity (H) using H = 1 - homozygosity

		p = 0.7 A	q = 0.3 B
0.7 (p)	A	$.7^2$ (p^2)	.7x.3 (p.q)
0.3 (q)	B	.7x.3 (p.q)	$.3^2$ (q^2)

genotypes:	AA		AB		BB	
frequency:	0.49	+	2 x 0.21	+	0.09	= 1.0
	p^2		2p.q		q^2	= 1.0
			(binomial expansion)			

Homozygosity $= p^2 + q^2$

Heterozygosity $= 2p.q$
or $= 1$ - homozygosity
$= 1 - (p^2 + q^2)$

or, if > 2 alleles $= 1 - (p_1^2 + p_2^2 + p_3^2 + \ldots p_n^2)$
$= 1 - \sum p_i^2$

Effects of population sub-division on heterozygosity

One of the main effects that population subdivision has on genetic diversity, is the reduction in **observed** H compared with **expected** H. This is shown in the following two examples:

Examples

- Mice (Hartl 1997 p. 112):

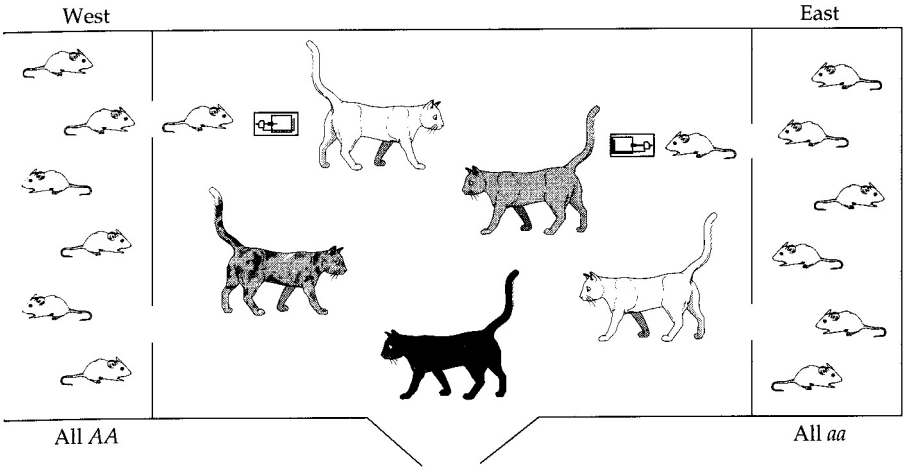


Figure 4.1 An extreme example of the general principle that a difference in allele frequency among subpopulations results in a deficiency of heterozygotes. The floor plan is that of a hypothetical barn. The mouse subpopulations in the east and west enclaves are completely isolated owing to the cats in the middle. The west subpopulation is fixed for the *A* allele and the east subpopulation for the *a* allele. Trapping mice at random in the area patrolled by the cats would yield an overall allele frequency of ½ but no heterozygotes.

- Blue & white flowers (Hartl 1997 p. 114)

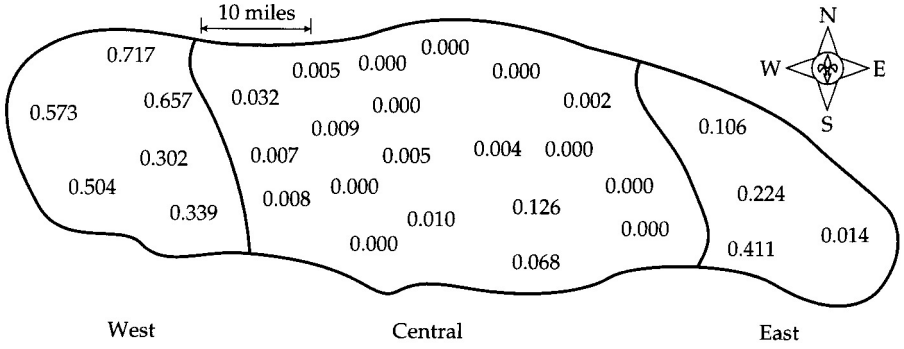


Figure 4.2 Estimated frequency of a recessive allele for blue flower color in populations of *Linanthus parryae* in an area of approximately 900 square miles in the Mohave desert. Each allele frequency is based on an examination of approximately 4000 plants over an area of about 30 square miles. (After Wright 1943a.)

TABLE 4.1 HIERARCHICAL STRUCTURE OF *LINANTHUS PARRYAE*

	Subpopulations		Regions		Total	
Region	Allele frequency	Heterozygosity	Average allele frequency	Heterozygosity	Average allele frequency	Heterozygosity
W	0.573	0.4893				
	0.717	0.4058				
	0.504	0.5000				
	0.657	0.4507				
	0.302	0.4216				
	0.339	0.4482	0.5153	0.4995		
C	9 × 0.000	0.0000				
	0.032	0.0620				
	0.007	0.0139				
	0.008	0.0159				
	0.005	0.0100				
	0.009	0.0178				
	0.005	0.0100				
	0.010	0.0198				
	0.068	0.1268				
	0.002	0.0040				
	0.004	0.0080				
	0.126	0.2202	0.0138	0.0272		
E	0.106	0.1895				
	0.224	0.3476				
	0.411	0.4842				
	0.014	0.0276	0.1888	0.3062	0.1374	0.2371
Average heterozygosity		$H_S = 0.1424$		$H_R = 0.1589$		$H_T = 0.2371$

Source: Data from Wright 1943a.

Derivation and explanation of F-statistics

The extent of reduction in observed H shown in these examples above can be used to quantify the level of genetic differentiation between the subpopulations. This quantification has been formalised (in the first instance by Wright (1951, 1965, 1978) in a series of hierarchical F-statistics.

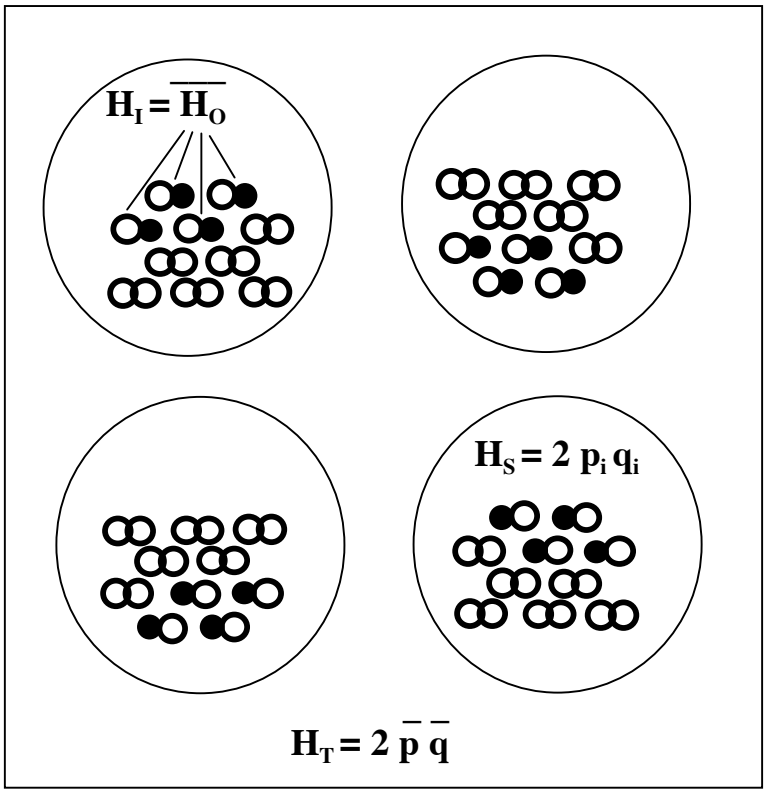
Firstly, a series of hierarchical measures of heterozygosity are defined:

H_I = mean <u>observed</u> heterozygosity per individual <u>within subpopulations</u>

H_S = mean <u>expected</u> heterozygosity <u>within</u> random mating subpopulations = $2p_iq_i$

H_T = <u>expected</u> heterozygosity in random mating <u>total population</u> = 2 p q

The quantities that these terms are measuring can be visualised as below:



Now, using these three different hierarchical measures of H, we can define three hierarchical F-statistics, defined below:

- **INBREEDING COEFFICIENT = $F_{IS} = (H_S - H_I) / H_S$**

- the mean reduction in H of an <u>individual</u> due to non-random mating within a <u>subpopulation</u>
- i.e., a measure of the extent of genetic inbreeding within subpopulations
- can range from –1.0 (all individuals heterozygous) to +1.0 (no observed heterozygotes)
- sometimes referred to simply as F rather than F_{IS}

- **FIXATION INDEX = $F_{ST} = (H_T - H_S) / H_T$**

- the mean reduction in H of a <u>subpopulation</u> (relative to the <u>total</u> population) due to genetic drift among subpopulations
- i.e., a measure of the extent of genetic differentiation among subpopulations
- can range from 0.0 (no differentiation) to 1.0 (complete differentiation – subpopulations fixed for different alleles)

- **OVERALL FIXATION INDEX = $F_{IT} = (H_T - H_I)/H_T$**

- the mean reduction in H of an individual relative to the total population

Note: F_{IT} combines contributions from non-random mating within demes (F_{IS}) and effects of random drift among demes (F_{ST}) –

The relationship between the three F-statistics is:

$$(1 - F_{IT}) = (1 - F_{IS})\,(1 - F_{ST})$$

Now, going back to our two initial examples in Hartl, we can see what F_{ST} means in a couple of real, simplified examples.

Mouse example:

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} \qquad \left(H_T = 2pq = 2 \times .5 \times .5 = 0.5\right)$$

$$= \frac{0.5 - 0.0}{0.5} = 1.0$$

that is, there is absolute differentiation between the 2 subpopulations., with each fixed for a different allele. Another way of thinking of this (obvious from fig 4.1) is that 100% of the total genetic variation is **between** subpopulations., with zero variation **within** subpopulations

Flower example (all subpopulations – not considering regions):

$$F_{ST} = \frac{0.2371 - 0.1424}{0.2371} = 0.39$$

That is, there is a substantial differentiation among all the subpopulations as can be seen from the great variation in allele frequencies in fig. 4.2. Putting this in other words, 39% of the total genetic variation is distributed **among** subpopulations, with 61% of the variation **within** subpopulations

Although F_{ST} has a theoretical range of 0 to 1.0, the observed maximum is usually much less than 1.0. (See below for the effect of highly variable loci such as microsatellites.) Wright (1978) suggests the following qualitative guidelines for the interpretation of F_{ST} (based on allozyme loci):

- the range **0.0 to 0.05** may be considered as indicating **little** genetic differentiation

- the range **0.05 to 0.15** indicates **moderate** genetic differentiation

- the range **0.15 to 0.25** indicates **great** genetic differentiation

- values of F_{ST} **above 0.25** indicate **very great** genetic differentiation

However, keep in mind that these are very general guidelines.

Table 3 (p.302) from Hartl (1989) gives some comparisons among F_{ST} values from a range of species to give some perspective as to what can be expected in natural populations.

***TABLE 5.1* ESTIMATES OF *Nm* AND $\hat{F}_{ST}$**

Species	*Type of organism*	*Estimated* Nm	*Estimated* $\hat{F}_{ST}$
Stephanomeria exigua	Annual plant	1.4	0.152
Mytilus edulis	Mullusc	42.0	0.006
Drosophila willistoni	Insect	9.9	0.025
Drosophila pseudoobscura	Insect	1.0	0.200
Chanos chanos	Fish	4.2	0.056
Hyla regilla	Frog	1.4	0.152
Plethodon ouachitae	Salamander	2.1	0.106
Plethodon cinereus	Salamander	0.22	0.532
Plethodon dorsalis	Salamander	0.10	0.714
Batrachoseps pacifica ssp. 1	Salamander	0.64	0.281
Batrachoseps pacifica ssp. 2	Salamander	0.20	0.556
Batrachoseps campi	Salamander	0.16	0.610
Lacerta melisellensis	Lizard	1.9	0.116
Peromyscus californicus	Mouse	2.2	0.102
Peromyscus polionotus	Mouse	0.31	0.446
Thomomys bottae	Gopher	0.86	0.225

Source: Data from Slatkin 1985.

Extension to hierarchical F-statistics

It can be seen conceptually, without too much difficulty, that these three F-statistics described above could be extended to include higher levels of hierarchy. For example, if we have a series of subpopulations which naturally fall into three separate groups (e.g., 3 river or ocean basins), we could add another level, groups, in our hierarchy to give: (1) variation among indivs. within subpopulations, (2) subpopulations, (3) groups of subpopulations, and (4) total variation. If we denote the group with subscript C (as done in Arlequin), this would give us the following F_{ST} –related statistics:

F_{ST} – the variance among subpopulations relative to the total variance
F_{SC} – the variance among subpopulations within groups
F_{CT} – the variance among groups relative to the total variance

The calculation of these values is a relatively straightforward extension of those shown previously. This hierarchy could be extended upwards further, if suitable for the data set. It could also be extended downwards, to consider variation within individuals, if you have diploid genotypic data.

Modifications to F_{ST} calculations required

The basic calculation formulae for F-statistics given above were derived as simple theoretical predictions. However, as with all statistics, we are trying to estimate a theoretical quantity using limited and imperfect data. That is, all we can actually do is to calculate an estimate, F_{ST}-hat, which hopefully closely approximates the true F_{ST} . To achieve as close an approximation as possible, we need to allow for a number of factors, listed below.

Multiple alleles and loci

The original formulation of F_{ST} by Wright considered only one biallelic locus. This was extended to first accommodate multiple alleles, and then to accommodate multiple loci (Wright 1978, Nei 1973). (The multiple locus version was termed G_{ST} by Nei, but I recommend sticking to F_{ST} for consistency). The formulae I have presented above are the multiple-allele forms. The multiple locus forms may be calculated in a number of ways, the simplest being to just average F_{ST} over loci, although it is more appropriate to average the H_S and H_T over loci. It seems an appropriate place here to emphasise how important it can be to estimate F_{ST} using as many loci as possible. Each locus is like a totally independent random trial. Among a group of subpopulations, each locus may, purely by chance, drift in frequency quite differently from the next locus. As a result, each locus will provide a different estimate of F_{ST} , sometimes radically so. Examples of this are shown in the tables in Box F (pp300-301, Hartl 1989). It is clear that an average over many loci will give a much better measure of differentiation among populations than a calculation from only one locus, such as the mt locus.

Locus	F_{IS}	F_{IT}	F_{ST}
ACP	0.0457	0.0627	0.0177*
ADA	0.0746	0.1013*	0.0289**
GPT	0.0514	0.0920*	0.0428**
HPT	−0.0420	−0.0230	0.0183
GC	0.0529	0.0659	0.0137
All loci	0.0392	0.0628*	0.0225**

Sampling effects

There also need to be adjustments made to these statistical calculations to account for the errors introduced by limited sampling. The first level of sampling to be considered is the limited sampling of individuals within a subpopulation Sometimes it is also necessary to account for the limited sampling of subpopulations from the total number of subpopulations within the species. This is necessary when you are making inferences about differentiation among all the subpopulations (a random effects model) (e.g., “This study therefore implies that there exists significant population structure among all subpopulations of species X”). This random effects model is usually used in ANOVA / AMOVA approaches (see below). However, this sampling correction is not strictly necessary if you are making inferences about only those populations you have sampled (a fixed effects model) (e.g., “This study therefore implies that there exists significant population structure among subpopulations A, B & C of species X”). The specific mechanisms of incorporating sampling error into your calculations won’t be dealt with here, because they become quite complicated, but it is important that you know whether the program you are using to calculate F_{ST} ‘s does incorporate such corrections (and what they are).

Haploid Data

The explanation given above using heterozygosity for determining F_{ST} for diploid loci is all well and good, but how do you define F_{ST} for a haploid locus, where heterozygosity is relatively meaningless? This has been approached in a number of ways (as usual), but the simplest way of reconceptualising F_{ST} for haploid loci is to think in terms of haplotype diversity instead of heterozygosity. Haplotype diversity (conveniently, also referred to as H) is a measure of the degree of variation in haplotypes found within a population, and is calculated as:

$$H = 1 - \sum_{i=1}^{j} p_i^2$$

This just happens to be the same way we can calculate heterozygosity in a diploid locus, although there are no heterozygotes here. To give you a quick conceptual idea of what haplotype diversity means in a real example, here is a worked example of two subpopulations, where the second subpopulation is

intuitively less diverse in haplotypes than the first. (I use the terms 'haplotype' and 'allele' interchangeably).

	Pop. 1		Pop. 2	
Haplotype	p_i	p_i^2	p_i	p_i^2
1	.5	.25	.9	.81
2	.4	.16	.1	.01
3	.1	.01		
		Σp_i^2=.42		Σp_i^2=..82
H		1-Σp_i^2=**0.58**		1-Σp_i^2=**0.18**

So, to now calculate F_{ST} for a haploid locus, just replace the H (heterozygosity) terms with the equivalent H (haplotype diversity) terms in the original formula above.

DNA Sequence Data

So far, we have been concerned only with allele (or haplotype) **frequencies** when calculating F-statistics. This is fine for allozyme and microsatellite data, where this is all the information we have. However, when we also have DNA sequence (or RFLP) data, we can determine how different each haplotype (or allele) is from each other, which we know gives us a lot more information about population substructure than we get from purely the haplotype frequencies. How can we incorporate this additional information into a similar measure of subpopulation differentiation?

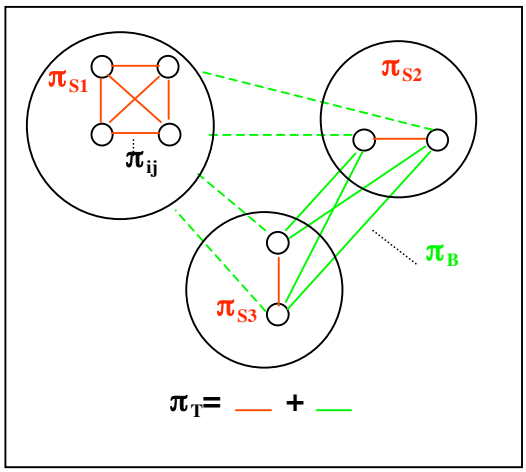
Slightly different measures have been devised by different authors (as usual!), but the simplest of these to comprehend in light of our previous discussion is probably that of Nei (1982). What he did was to define a similar measure of population differentiation as F_{ST} , but this time using a measure of **nucleotide diversity** ($\boldsymbol{\pi}$) within a population, in place of heterozygosity (H) or haplotype diversity (H). If we define π_{ij} as the genetic distance between haplotype i and haplotype j (measured either by the simple proportion of nucleotide differences, or by some more complicated method, e.g., Jukes-Cantor, Kimura 2-parameter, etc.), then the nucleotide diversity within the **total** population is

$$\pi_T = \sum_{ij} p_i p_j \pi_{ij}$$

where p_I and p_j are the overall frequencies of haplotypes i and j respectively. That is, the distances between haplotype pairs are simply weighted by how common they are, to arrive at an average. If we also define π_S as the average nucleotide diversity within **subpopulations**, then we can derive a familiar expression for an F_{ST} –like nucleotide measure of subpopulation differentiation:

$$\frac{\pi_T - \bar{\pi}_S}{\pi_T} = \frac{\pi_B}{\pi_T}$$

Here, π_B is the average nucleotide diversity **between** subpopulations What all these terms refer to conceptually can be seen in the following figure:



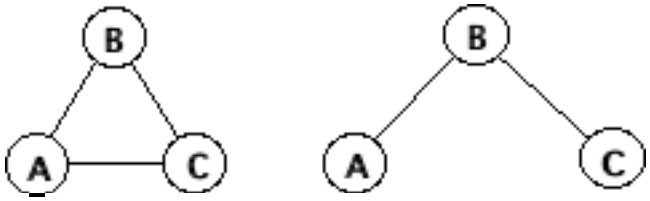
$\bar{\pi}_S=(\pi_{S1}+\pi_{S2}+\pi_{S3})/3$

$\pi_T - \pi_S = \pi_B$

This statistic could also be called F_{ST} , but it was originally described by Nei (1982) as γ_{ST} . A related statistic derived by Lynch & Crease (1990) was called N_{ST} , one derived for mtDNA data by Takahata and Palumbi (1985), G_{ST} , and one by Excoffier et al. (1992, see below), Φ_{ST} (phi-st). Although each of these statistics for nucleotide data is calculated slightly differently, in reality they are all trying to estimate the same parameter - the proportion of nucleotide diversity among subpopulations, relative to the total – and their values are usually quite similar, particularly with larger sample sizes. The same things can be said for all the different ways of calculating F_{ST} from allelic data (including G_{ST} and θ_{ST}). Given the multitude of different descriptor variables used by different authors, and the fact that within the two classes they are trying to estimate essentially the same parameters, my convention is to refer to the allelic form of the statistic as F_{ST} , and the nucleotide diversity form as Φ_{ST} , and simply mention somewhere whose formulae or program you used to calculate them.

There is also a very simple conceptual relationship between the allelic F_{ST} and the nucleotide Φ_{ST} , shown below. In the allelic calculations (F_{ST}), it is assumed that all alleles are equidistant from each other, while in the nucleotide diversity calculations (Φ_{ST}), there are different distances between different alleles. (This can indeed be applied to any other type of distance calculation you might require, such as between microsatellite alleles). In fact, now we can actually calculate the allelic F_{ST} in exactly the same way we calculate Φ_{ST} (in AMOVA – see below) by simply making all the distances between alleles equal one (i.e., replace the pairwise distance matrix by a unity matrix). This is shown below.

F_{ST} **Φ_{ST}-**



Pairwise Distances (π_{ij})

	A	B	C
A			
B	1		
C	1	1	

	A	B	C
A			
B	1		
C	2	1	

Below, I also provide a worked example of a simple data set, using the allele relationships shown above, showing the difference in values of F_{ST} from the allelic form of the calculation and the nucleotide form of the calculation.

Alleles	Pop. 1 Freq.	Pop. 1 Rel. Freq.	Pop. 2 Freq.	Pop. 2 Rel. Freq.	Total Freq.	Total Rel. Freq.
A	4	.8	0	0	4	.4
B	1	.2	1	.2	2	.2
C	0	0	4	.8	4	.4
$\underline{F_{ST}}$:						
Σp_i^2=		$.8^2+.2^2=.68$		$.8^2+.2^2=.68$		$.4^2+.2^2+.4^2=.36$
$H=1-\Sigma p_i^2$=		1-.68=.32		1-.68=.32		1-.36=.64
$\underline{\Phi_{ST}}$:						
$\Sigma p_i p_j \pi_{ij}$=		.8x.2x1=.16		.2x.8x1=.16		.4x.2x1=.08 +.2x.4x1=.08 +.4x.4x2=.32 **π_T=0.48**

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T}$$

$$= \frac{0.64 - 0.32}{0.64} = \frac{0.32}{0.64} = 0.5$$

and

$$\Phi_{ST} = \frac{\pi_T - \overline{\pi}_S}{\pi_T} = \frac{\pi_B}{\pi_T}$$

$$= \frac{0.48 - 0.16}{0.48} = \frac{0.32}{0.48} = 0.67$$

As you would expect from this data set, the Φ_{ST} value is greater, but this is not always the case, and the reverse is true of some data sets. As explained under gene flow, differences in value between the two forms of F_{ST} do not signify any problem or inconsistency, but are actually telling you something interesting about your data. The two forms are really measuring different properties of your data, and one is not necessarily any 'better' than the other. Even in terms of simply detecting whether or not there is significant differentiation between subpopulations , it depends entirely on your data set as to which form of F_{ST} - allelic or nucleotide distance – is the more powerful statistically. So the bottom line is that it is useful to calculate **both** types of F_{ST} for your data set.

Microsatellites and R_{ST}

The most recent adaptation of the F_{ST} statistic has been to microsatellite data. The standard allele-frequency type of F_{ST} calculation is certainly suitable for application to this data (keeping in mind the potential limits to the maximum value of F_{ST} in highly variable loci, mentioned below). However, **if** microsatellite loci evolve in accordance with a stepwise mutation model (SMM) (Valdes et al., 1993), then microsatellite alleles that are of similar length (i.e., have similar no. of repeats) are likely to be more closely related to each other than alleles that are very different in length (i.e., have very different no. of repeats). That is, microsatellite data **may** give us good information about the relationships among alleles, unlike allozyme data. In the previous section I showed that, by using a Φ_{ST} –type statistic, we can use the additional information that DNA sequence data provides about the relationships among alleles, to thence

provide more information about the relationships among subpopulations In a very analogous way, if we can derive an F_{ST} -type statistic that uses the additional information that microsatellite data may provide about the relationships among alleles, then we can obtain more information about the relationships among subpopulations Several authors have done this (e.g., Slatkin, 1995, Goodman, 1997) (once again, each using slightly different calculations) and called this new statistic R_{ST} (or ρ). Slatkin defined the statistic in terms familiar to us:

$$R_{ST} = \frac{S_T - S_W}{S_T}$$

where S_T (actually termed S-bar by Slatkin) is the variance in allele size in the total population, and S_W is the variance in allele size within subpopulations. Mikalakis and Excoffier (1996) and Rousset (1996) have shown that this term can also be similarly estimated by Φ_{ST} in an AMOVA framework (see below), after calculating appropriate distances between alleles (based on the sum of the squared differences in repeat length).

The verdict is still out on whether this R_{ST} statistic is actually better than F_{ST} for microsatellites (i.e., gives a better picture of reality). My view is that all microsatellite loci do not evolve in the same way. That is, some loci may fit the SMM, while others do not. Also, the longer the time of divergence between subpopulations, the greater the chance that alleles of the same length have evolved from different ancestors (i.e., are not identical-by-descent, or, are homoplasious), or have not evolved stepwise, and thus an SMM model may be misleading. In short, F_{ST} makes less assumptions, and thus is a more conservative measure. You can always calculate R_{ST} as well. If this gives you very similar results, then there is no problem. If it gives you different results, then this raises various hypotheses about the evolution of your microsatellite loci, which you may need to look into in more detail. For example, it may be obvious from the distributions of allele size that some or all your loci do not fit the SMM.

Alternative Derivations

The descriptions and derivations of the various F-statistics presented so far have come from just one of the many ways in which F_{ST} can be conceptualised. Various authors have derived F_{ST} in terms of:

- The correlation between uniting gametes, or the variance in allele frequency (V_q) (Wright's original derivation, 1951). Here,
  $F_{ST} = V_q / pq$

- The loss of heterozygosity over time in subpopulations with no migration
  $F_{ST} = 1 - (1 - 1/2N)^t$

- The probabilities of identity-by-descent:
  $F_{ST} = (f_0 - \mathbf{f}) / (1 - \mathbf{f})$,
  where: f_0 = Pr. identity of two alleles from same popn.,
  $\mathbf{f}$ = Pr. identity from total popn.

- The average times to coalescence of subpopulations (Slatkin 1991):
  $F_{ST} = (\mathbf{t} - t_o) / \mathbf{t}$,
  where $\mathbf{t}$ = average time to coalescence for total population
  t_0 = average time to coalescence within a population

- An ANOVA-like approach to partitioning genetic variance into within- and among-subpopulation components (Weir and Cockerham, 1984).

Apart from the last of these approaches, I won't go into any detail of these different concepts of F_{ST} here. It is sufficient at this point to realise that they exist, at the very least so that you aren't too confused when you come across them. I'll deal in more depth here with the ANOVA approach, as it has lead to the AMOVA approach, which is now commonly used for DNA sequence data sets.

ANOVA

It may be useful here to very quickly revise the broad concepts behind the usual ANOVA procedure, as they are much easier to conceptualise with real numbers than with gene frequencies.

So, a quick, related example of ANOVA on real numbers: comparing the mean length of whales among subpopulations If we have *i* subpopulations and *j* individuals in each subpopulation, then our model for partitioning variation in length among individuals is:

$$x_{ij} = X + a_I + b_{ij}$$

where	x_{ij}	is the length of an individual	with variance	σ^2
	X	is the overall mean length		
	a_I	is the subpopulation effect	with variance	σ_a^2
and	b_{ij}	is the indiv. variation effect	with variance	σ_b^2

We test the significance of the variance among subpopulation mean lengths by calculating the ratio of the subpopulation variance to the total variance

$$\mathbf{F} = \sigma_a^2 / \sigma^2$$

(Although apparently similar to F_{ST} , this more familiar F is actually quite different, and can have values greater than 1.0, unlike F_{ST}).

Our example data set of whale lengths:

Subpopulation 1: Indiv. A: 4m, B: 5m, C: 6m mean = 5
Subpopulation 2: Indiv. D: 8m, E: 9m, F: 10m mean = 9

Overall mean (X) = 7

And following our model of partitioning variance, we can see the variation among individuals' lengths as being composed of the overall mean plus the subpopulation and the indiv. effects:

x_{ij}		=	X	+	a_I	+	b_{ij}
x_{1A}	= 4	=	7	+	- 2	+	- 1
.							
.							
.							
x_{2F}	= 10	=	7	+	2	+	1
	σ^2				σ_a^2		σ_b^2 - related variances

Now, if we think about allele frequency variance instead of length variance, we can use a very similar model (derived by Weir & Cockerham 1984):

$$x_{ij} = X + a_I + b_{ij}$$

where x_{ij} is the probability of an individual having allele A with variance σ^2
X is the overall probability of having allele A
a_I is the subpopulation effect with variance σ_a^2
and b_{ij} is the indiv. variation effect with variance σ_b^2

Once again, we measure the size of the variance among subpopulation allele frequencies by calculating the ratio of the subpopulation variance to the total variance

$$\mathbf{F_{ST}} = \sigma_a^2 / \sigma^2$$

This term was called θ by Weir and Cockerham (1984) (just one more variant to keep us on our toes, I suspect). However, in this case we cannot realistically assume any particular parametric distribution of this statistic in order to test its significance, and usually resort to non-parametric randomisation tests (see below).

AMOVA

In a similar way that F_{ST} was adapted to γ_{ST} by incorporating a measure of nucleotide distance between alleles, so too θ was adapted to Φ_{ST} by incorporating a measure of distance between haplotypes (Excoffier et al. 1992). A major benefit of the Φ_{ST} approach is that it is very flexible, and **any** type of distance you think is appropriate can be used - hence its extension to microsatellite data also (described above).

I won't go into too much detail here about the AMOVA approach, except to say that this is the approach used in the ARLEQUIN program, and it is excellently documented in its manual (which I thoroughly recommend reading if you use the program).

Comparing F_{ST} 's

The F_{ST} statistic has proved very valuable in many ways, despite some limitations. It is a very convenient and common way to compare differentiation among two or more subpopulations It immediately gives a pretty good idea of the degree of subpopulation structure in any organism, without knowing anything else at all. It also has a series of good statistical properties. These include (1) its potential relationship with gene flow (see below), and (2) its relatively rapid approach to equilibrium - i.e., it approaches equilibrium a lot faster than many other measures (see fig 17 on p315 of Hartl 1989), but it can still take a long time – on the order of $1/m$ generations! (where m is the migration rate).

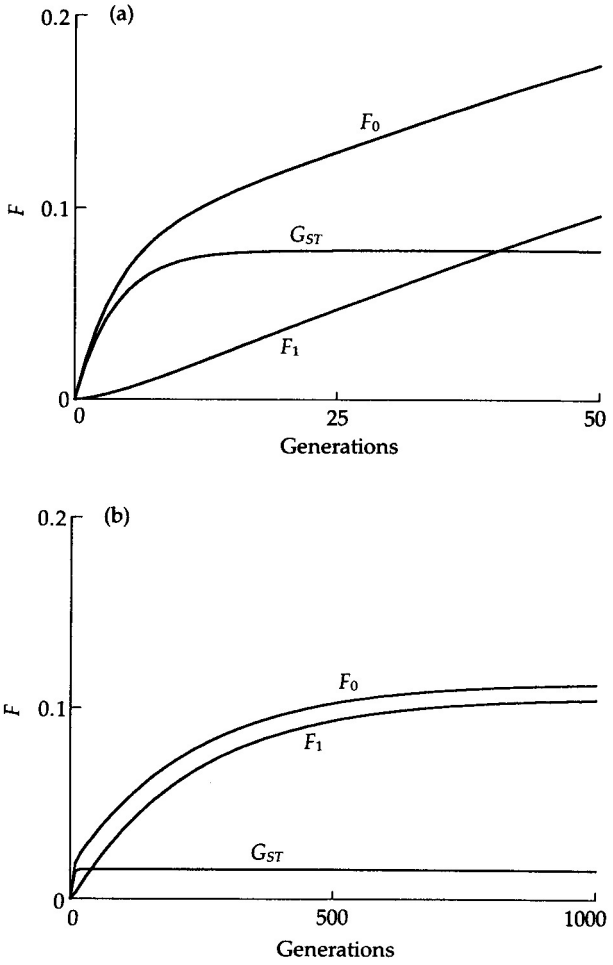


Figure 17. The changes in F_0, F_1, and G_{ST} in the n-island model according to Equations 6.28. For both (a) and (b) the number of individuals per island was $N = 20$, the number of islands was $n = 10$, the mutation rate was $\mu = 0.001$, and the migration rate was 0.1. (a) The first 50 generations of the simulation, starting with all identity measures equal to zero, and demonstrating how rapidly G_{ST} attains equilibrium. (b) The figure verifies that the other identity measures also attain a steady-state, but after a much longer time.

Perhaps one of its best attributes is that F_{ST} can be a very simple descriptor of the degree of population substructure, for comparison among loci or species. However, this can be problematic if the level of variation (or heterozygosity) within subpopulations varies dramatically among loci or species. F_{ST} is a **relative** rather than an **absolute** measure of differentiation. Therefore, if some loci (or species) have very high levels of variation compared to others, then a simple comparison of F_{ST} 's would be misleading. The extreme example of this bias occurs when dealing with microsatellite loci, which often have very high heterozygosities, sometimes approaching 1.0. Various authors have pointed out that, for a given absolute level of subpop differentiation, highly variable loci will, by definition, give a lower value of F_{ST} than less variable loci. As Hedrick (1999) showed,

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} = 1 - \frac{\overline{H}_S}{H_T} \; < \; 1 - \overline{H}_S = \text{homozygosity}$$

In other words, although the theoretical maximum value for F_{ST} is 1.0 (when each subpop is fixed for a different allele), the practical **upper limit** for F_{ST} is actually the level of (expected) homozygosity. For many microsatellite loci this means that F_{ST} must be smaller than perhaps 0.1 or less.

Another problem comparing F_{ST} –like values among loci occurs with the mitochondrial locus. Because it is haploid and (usually) only maternally inherited, its effective population size is 1/4 that of a nuclear locus. Thus mtDNA would be expected (on average) to diverge between subpopulations four times as

quickly as a nuclear locus, and hence F_{ST} values calculated from mtDNA may be four times larger than those from nuclear data from the same populations.

The bottom line of all this discussion is that F_{ST} values can be extremely useful for comparing loci or, more particularly, species, but that first you have to be sure you are comparing apples with apples (or should that be kiwi fruit with kiwi fruit?).

Calculating probability and testing significance

Firstly, we need to be clear exactly what we want to test here. Usually, we want to know if there is significant population differentiation among our samples. Hence our null hypothesis is that $F_{ST} = 0$, and our alternative is that $F_{ST} > 0$. This need not necessarily be the case, however, as you may possibly be interested in comparing F_{ST}'s between groups of subpopulations or species, and therefore your H_1 in this case may be, for example, that $F_{ST} < 0.2$

The simplest (and earliest) method of calculating the probability of an F_{ST} value was by using the simple equation $\chi^2 = 2N F_{ST}$, and comparing this value to the standard χ^2 distribution. (It requires some modifications for multiple alleles and loci). However, this requires us to make a big assumption about the distribution of F_{ST} values, which is probably not valid. A very useful non-parametric approach is to jackknife or bootstrap over loci, which provides approximate confidence intervals, and is still employed by some programs. However, this is pretty useless unless you have something like six or more loci, and of course is impossible if you have only one. The most flexible approach, and which is used most commonly today, is to randomly permute (for example) individuals among the subpopulations, calculate an F_{ST} value, and repeat this 100-1000 times to give a 'random' distribution of the statistic, against which you compare your 'true' statistic. The random probability of getting your F_{ST} (or greater) is then simply the proportion of the randomised values that are equal to or greater than your 'true' value. Using this process you can also test hypotheses other than $F_{ST} = 0$, but you need to look at the actual distribution of randomised values yourself (you can do this in ARLEQUIN).

One last, but very important, point about calculating probabilities in the context of subpop differentiation (or generally): When a number of tests are performed at the same time, for example in a matrix of pairwise F_{ST}'s between subpopulations, then the probabilities actually should be adjusted for the fact that one of the many tests may be significant simply by chance. A Bonferroni-type adjustment should be applied to account for this (e.g., Rice 1989). My understanding is that ARLEQUIN does **not** do this automatically, so you need to account for this yourself in interpreting your results. If none of this makes any sense to you, please come and ask me, or look at a general stats book!

What F_{ST} may tell us about population divergence and gene flow

F_{ST} has proved to be a very useful parameter in many respects, as described above. One major advantage is the possibility that it may tell us a lot about the **processes** leading to divergence between subpopulations or the maintenance of that divergence. When two subpopulations begin to diverge after, say, a vicariant event that separates them (e.g., a mountain range uplifts), then two processes begin acting, and in opposite directions. The first process is that, under the influence of genetic drift (see fig. above), the subpopulations start to diverge genetically. Over time, as seen in fig. 7.11 (p 286 Hartl 1997), F_{ST} will gradually increase, until it finally approaches 1.0, **if** there is no continued migration between the subpopulations.

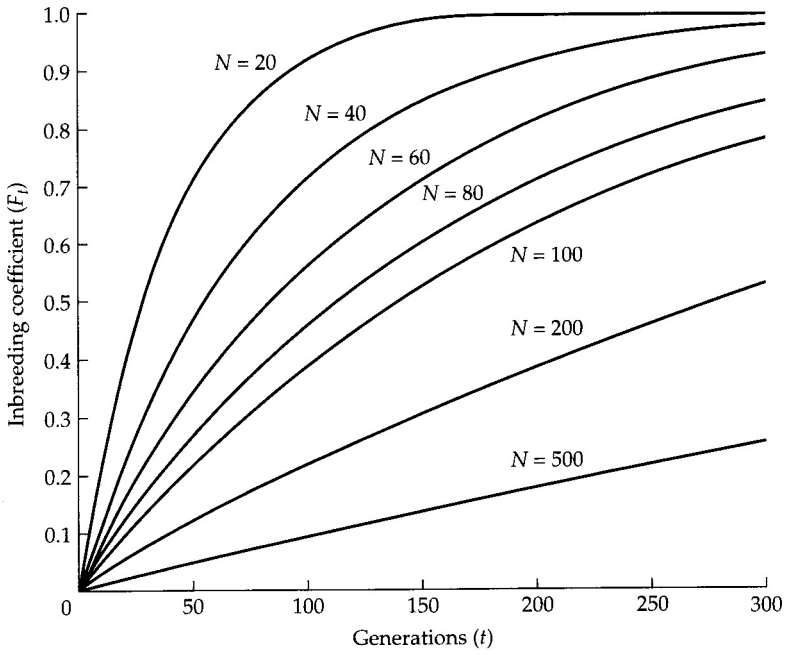
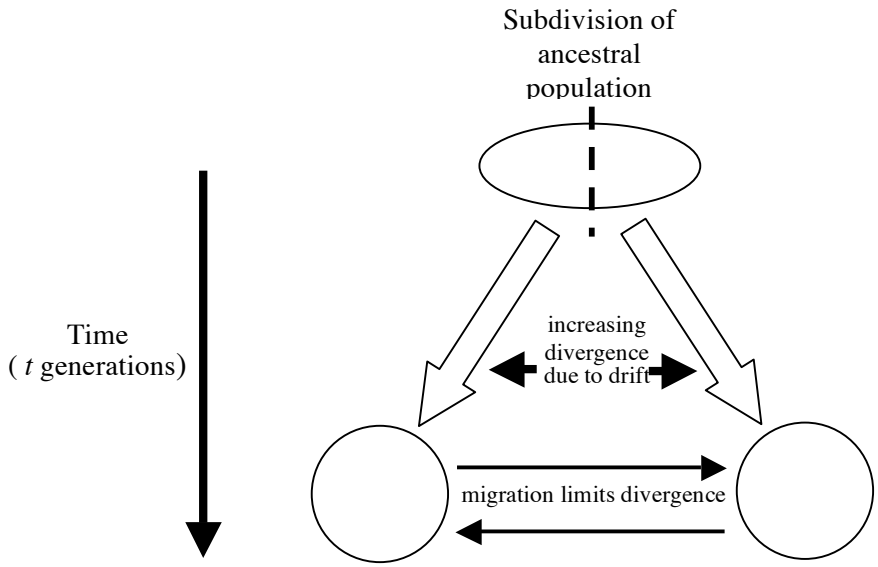


Figure 7.11 Increase of F_t in ideal populations as a function of time and effective population size N.

On the other hand, there is likely to be some level of continued migration between the two subpopulations, as they diverge. This migration will tend to limit the genetic divergence between the subpopulations Thus there are two opposing forces determining the divergence between subpopulations, and hence, F_{ST} . Genetic drift over time will allow them to diverge, while migration acts to keep them similar (see fig below).



When we determine the level of genetic divergence between two subpopulations at a point in time, i.e., F_{ST} , we know that the value is due to one or both of these two effects, i.e., drift over time, or migration. **If** we can assume that divergence is due entirely to one or the other effect, then we can use the value of F_{ST} to estimate either 1) time since subpop divergence, or 2) migration. However, we cannot estimate both. **Either**, 1) we assume that there has been no migration between subpopulations since divergence, in order to estimate time since subpop divergence, **or** 2) we assume that divergence and F_{ST} have reached equilibrium, in order to estimate migration. If we can make neither assumption, then we simply know that

the present level of divergence (F_{ST}) is due to some combination of both effects, and therefore we can't estimate either parameter.

Considering the cases where we can make one of these assumptions, we will now look at the implications in more detail.

F_{ST} and time since divergence

Given our first scenario above (that a single population splits into two subpopulations at some point, which then each diverge randomly over time with no migration), then it is relatively easy (!) to come up with an equation showing how F_{ST} increases over time. (For simplicity, here we will refer to F_{ST} as simply *F*, and use the subscript *t* to indicate time in generations, i.e., F_t is the value of F_{ST} after *t* generations.). This equation is

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t$$

Fig. 7.11 (p 286 Hartl 1997) shows this relationship for various values of N. (It can be seen that for low values of N, F rapidly approaches the limit of 1.0. In general, F goes halfway to equilibrium in 1.39N generations). Therefore, if we know *F*, and *N* (actually N_e) then we can calculate *t* (in generations). Sounds lovely doesn't it? The only problems are that we usually have only fairly rough estimates of F and Ne from our data, and that we must assume that our subpopulations are behaving in an ideal manner, let alone our initial assumption of no migration!

However, all is not lost. We may have only a very rough estimate of the *absolute* time since divergence for two subpopulations, but it does give us a better way to estimate *relative* divergence between a series of pairs of subpopulations We see from fig 7.11 that the relationship between *F* and *t* is very non-linear. If we transform F to make it much more linear, then we will have a much better measure of divergence between a pair of subpopulations This is what has been done in deriving Reynolds' distance and Slatkin's linearised F_{ST} (calculated in ARLEQUIN). If all the assumptions are correct, then both of these give values in *t/N* generations, and are proportional to the divergence time. Such distances may be valuable for constructing trees or other graphical patterns (such as multidimensional scaling) of the genetic relationships among subpopulations (Other useful measures here are Nei's total nucleotide diversity between subpopulations (D_{XY}) and net nucleotide diversity (D_A) between subpopulations)

F_{ST} and gene flow

Now for our second scenario from above: that a single population splits into subpopulations at some point, which then each diverge randomly over time, but in this case migration has limited the extent of divergence, and the subpopulations have reached this limit and are at equilibrium. Now, given these assumptions, along with the assumption that all subpopulations can share migrants with all other subpopulations with equal chance (i.e., an **island model** of population structure), then there is an extraordinarily simple relationship between F_{ST} and migration (*m)*, which at its simplest is expressed as:

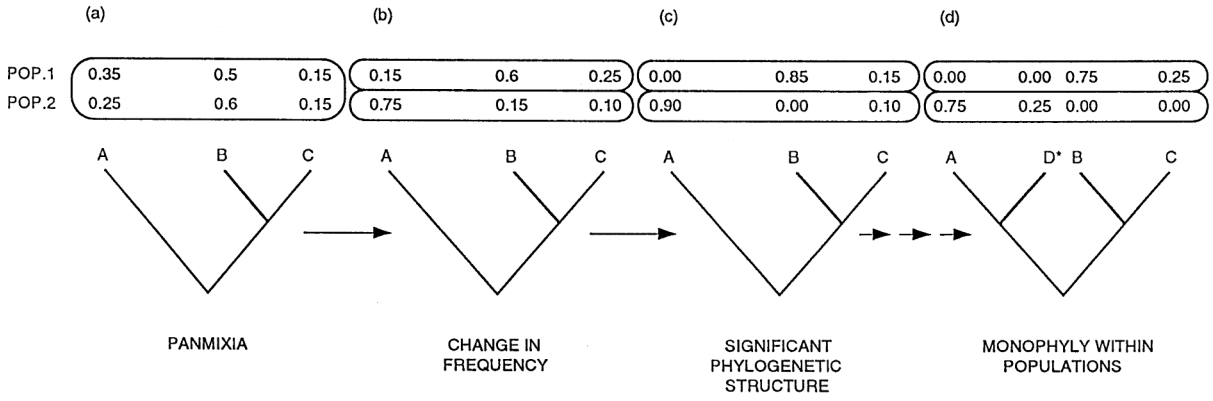
$$F_{ST} = \frac{1}{4Nm + 1}$$

(The relationship between F_{ST} and Nm is shown in fig.15, p 312 Hartl 1989). The extraordinary thing here is that the terms have simplified so that *N* and *m* appear together in one term *Nm* (given more assumptions, such as *m* is very small, and the mutation rate, μ, is much smaller than *m*). This is very satisfying (for a nerd), because 1) we no longer have to worry about trying to estimate that ever-tricky parameter N_e ,and 2) the term *Nm* is actually something meaningful. Think about it – the product of the effective population size (N_e) and the proportion of the population that migrates (*m*) is the *actual* number of individuals that migrate (per generation). Pretty wild isn't it? Everyone else thinks so too, which is why this value *Nm* (or just *M*, as referred to by Slatkin, and ARLEQUIN) has been so used and abused over

the years. The problem is of course that, given all the assumptions, it is pretty courageous to interpret *Nm* as the true number of individuals migrating. However, as before, it is a very valuable *relative* measure of migration between subpopulations Furthermore, it can also be compared with alternate estimates of *Nm* that have been derived independently from the data (e.g., from coalescence approaches). All in all, it is a pretty useful relationship, but is ultimately just a different way of expressing the same information (F_{ST}).

F_{ST} and Φ_{ST} over time: how allele frequency differences and nucleotide diversity differences are not the same

To finish this brief discussion of 'Fun with F_{ST} 's', I think it's worthwhile to highlight a potentially useful conceptual difference between the F_{ST} and Φ_{ST} parameters. Briefly, it relies on the simple concept that allele frequencies can change quite rapidly (over only a few generations, if N_e is small), while complete fixation of alleles takes longer, and for new alleles to arise through mutation probably longer again (when mutation rates are relatively low). This simple series of events, therefore, is likely to take place in this order in subpopulations, as shown diagramatically below.



What this ultimately can mean is that, after a population splits, until subpopulations have reached a stable equilibrium (& for many organisms this is likely to not be the case at present), then F_{ST} is likely to increase first, and only after new alleles have arisen, and monophyletic clades of alleles have begun to arise in different subpopulations, will Φ_{ST} begin to increase substantially. That is, F_{ST} may be an indicator of short-term or recent population processes, while Φ_{ST} may be an indicator of longer-term or older processes.

References

Crow JF, Aoki K (1984) Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Science (USA)*, 81: 6073-6077.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131: 479-491.

Goldstein DB, Linares AR, Cavallisforza LL, Feldman MW (1995) An Evaluation of Genetic Distances For Use With Microsatellite Loci. *Genetics*, 139: 463-471.

Goodman SJ (1997) R-ST Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology*, 6: 881-885.

Hartl D, Clark A (1989) *Principles of Population Genetics. 2nd Ed.* Sinauer, Sunderland MA.

Hartl D, Clark A (1997) *Principles of Population Genetics. 3rd Ed.* Sinauer, Sunderland MA.

Hedrick PW (1999) Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution*, 53: 313-318.

Hedrick PW (2000) *Genetics of Populations*. Jones and Bartlett, Sudbury, Mass.

Lynch M, Crease TJ (1990) The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution*, 7: 377-394.

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, 142: 1061-1064.
Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70: 3321-3.
Nei M (1982) *Evolution of human races at gene level*. In: Human Genetics, Part A: The Unfolding Genome Bonne-Tamir B (Ed.) 167-181. Alan R. Liss Inc., New York.
Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, 22: 201-204.
Rice WR (1989) Analysing tables of statistical tests. *Evolution*, 43: 223-225.
Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, 142: 1357-1362.
Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research, Cambridge.*, 58: 167-175.
Slatkin M (1995) A Measure of Population Subdivision Based On Microsatellite Allele Frequencies. *Genetics*, 139: 457-462.
Takahata N, Palumbi SR (1985) Extranuclear differentiation and gene flow in the finite island model. *Genetics.*, 109: 441-457.
Valdes AM, Slatkin M, Freimer NB (1993) Allele Frequencies At Microsatellite Loci - the Stepwise Mutation Model Revisited. *Genetics*, 133: 737-749.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, 38: 1358-1370.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, 15: 323-354.
Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution.*, 19: 395-420.
Wright S (1978) *Evolution and the Genetics of Populations. Vol. 4. Variability Within and Among Natural Populations.* Univ. of Chicago Press, Chicago.