Introduction to Population Genetics

Notes by Rachel Fewster, *r.fewster@auckland.ac.nz*
Department of Statistics, The University of Auckland, New Zealand

Our aim here is to provide a coherent introduction to the basic concepts of Population Genetics. Population genetics is a difficult subject for self-study, especially from a statistical perspective. In the literature, you will see multiple different definitions claiming to be the same quantity; multiple different quantities that seem to have the same definition; and many confusions and vagaries regarding the distinction between parameters and statistics, the distributions underlying means and variances, and so on. Our aim is to provide a clear description of basic principles, and link together the common ways of describing and conceptualising key quantities.

We use the following text as our key reference:
Weir, B.S. (1996). *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.

A model for reproduction

The setting for all our models is as follows. Although these models are not biologically realistic, they are very helpful for understanding basic genetic processes.

- Diploid organisms: this means that every individual has two alleles at every locus.
- Monoecious: this means we don't distinguish between individuals of different sexes. All individuals can create offspring with all other individuals, including themselves.
- Non-overlapping generations.
- Fixed population size of N individuals every generation. Because the organisms are diploid, this means that we have $2N$ alleles available in every generation.
- Infinite number of gametes. Statistically, this is the same as saying that gametes are sampled at random *with replacement* to make the next generation. Biologically, we can think of each of the $2N$ alleles emitting an infinite number of copies of itself, and these copies mixing together before $2N$ lucky ones are selected at random to form the next generation. The proportion of different allele types (A, B, etc) in the infinite mix remains the same as it was in the original $2N$ copies, so allele types that are common in the original generation of size $2N$ will likely remain common in the next generation, whereas rare allele types might fail to be sampled and become extinct.

We will use different colours to depict different **allele states** or **types**: for example, alleles of type A will be given the same colour regardless of which individual they appear in. Figure 1 shows an animation of the conceptual reproduction scheme.

Note: A *gamete* can be thought of as the offspring of a single allele. If there are multiple loci, we need to be more precise: a gamete is effectively half of a genotype, where the selection of which half is made at random for every locus. For example, if an individual has genotype AB at locus 1, and genotype GH at locus 2, it can produce any of the following two-locus gametes: AG, AH, BG, and BH. In humans, each sperm cell and each egg cell correspond to a single gamete. Two gametes unite to make an offspring.

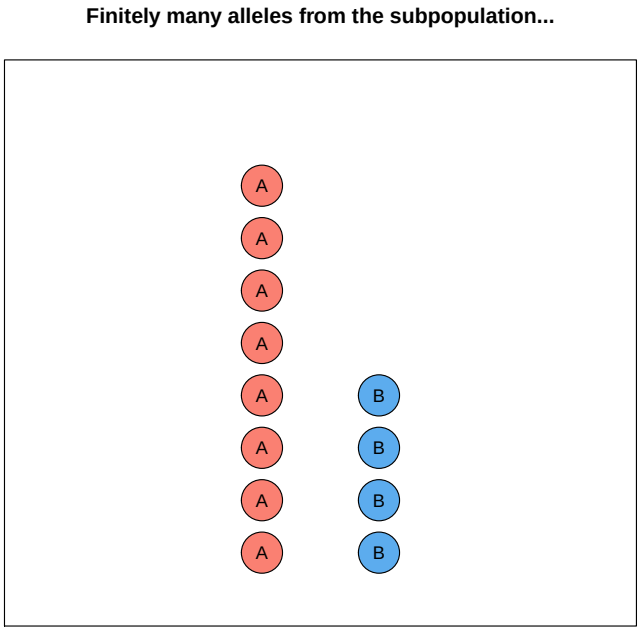


Figure 1: A finite number of individuals $N = 6$ with $2N = 12$ alleles, arranged into two allele types, A and B. At reproduction, each of the alleles spawns an infinite number of gametes, which mix together. To make the next generation, $2N$ of these gametes will be selected at random.

Hardy-Weinberg Equilibrium

Hardy-Weinberg Equilibrium, or HWE, is a long name for a simple phenomenon. Consider the infinite number of gametes at the end of the animation in Figure 1. The proportions of allele types A and B are $p_A = 8/12$, and $p_B = 4/12$. To make a new individual for the next generation, we sample two of these gametes and 'unite' them, like a sperm and an egg. The genotype of the new individual can be AA, AB, or BB. Note that we can't distinguish between genotypes AB and BA, so they are written as the single genotype AB.

If the two gametes are sampled independently, the probabilities of the different genotypes are:

$$\left.\begin{array}{rcl} \mathbb{P}(AA) & = & p_A^2 \\ \mathbb{P}(AB) & = & 2\,p_A\,p_B \\ \mathbb{P}(BB) & = & p_B^2 \end{array}\right\} \text{Hardy-Weinberg proportions.}$$

Thus *Hardy-Weinberg proportions* describe the genotype probabilities obtained if an individual's two gametes are sampled independently from the infinite gamete pool. This way of producing offspring is called *random mating*.

If a population is very large ($N \to \infty$), and gametes are united independently as above, then:

- Allele frequencies will not change from one generation to the next: i.e. $p_A(t+1) = p_A(t)$ where $p_A(t)$ denotes the frequency of allele type A at generation t;
- Genotype frequencies will follow the Hardy-Weinberg proportions, and therefore they will not change either.
- The population is said to be in *Hardy-Weinberg Equilibrium*, with the word 'equilibrium' meaning that the allele and genotype frequencies are constant over the generations.

The *Hardy-Weinberg Principle* states that a very large population ($N \to \infty$) will be in Hardy-Weinberg equilibrium *unless* some specific disturbing force is acting. Possible disturbing forces include:

- Non-random mating: e.g. due to inbreeding. Inbreeding occurs when gametes are more likely to unite with other gametes of their own type than they would under random mating, and it can arise from mating of close kin.
- Mutation or migration: these introduce new alleles into the population and disturb the binomial proportions.
- Selection: this implies some allele types or genotypes are more likely to create offspring than others, because they convey some advantage to survival.

If, by contrast, N is small, but mating is still random, the population may follow Hardy-Weinberg proportions every generation. However, the allele frequencies are likely to change each generation due to random sampling, and therefore the population is not strictly in Hardy-Weinberg Equilibrium. The change in allele frequencies from one generation to the next is called ***genetic drift***.

By illustration, recall the example in Figure 1 with $N = 6$. At generation t we have $8/12$ alleles of type A, and $4/12$ alleles of type B. For generation $t+1$, we sample 12 new alleles at random and independently, each with $p_A(t) = 8/12$ and $p_B(t) = 4/12$. We might get 6 As and 6 Bs. This greatly changes the allele frequencies for generation $t+1$ to $p_A(t+1) = 6/12 = p_B(t+1)$. This population is tiny and is subject to severe genetic drift. A much larger population with the same allele frequencies can not drift far from one generation to the next, under random mating.

The Wright-Fisher Model

The Wright-Fisher model is the fundamental model of population genetics. It examines genetic processes for a single locus. The structure is as follows:

- Diploid organisms: 2 alleles for each individual at the locus in question.
- Monoecious: no sexes, and selfing allowed.
- Non-overlapping generations.
- Fixed population size of N individuals every generation.

- Infinitely many gametes.
- Random mating: alleles are drawn independently from the infinite gamete pool.
- No selection: no allele or allele type is favoured over any other when selecting $2N$ gametes to form the next generation.
- No mutation: alleles can not change into a different type, and new allele types can not enter the population.

The Wright-Fisher model looks at the genetic composition of a population at generations $t = 0, 1, 2, \ldots$, when it is kept in these conditions. Of interest are the allele types, but also the individual allele lineages: for example, how many of the alleles available at time t can be traced back directly to the same ancestor.

At this point we meet an idea of fundamental importance in population genetics: the **infinite reference population** or **infinite ancestral population**. This is a population of infinite size, in Hardy-Weinberg equilibrium, from which **Generation 0** of the Wright-Fisher model will be drawn. The Wright-Fisher process has a fixed and finite size of $2N$ alleles from generation 0 onwards, but it is drawn from an infinite ancestral pool. We might be interested in examining the current generation to deduce how long ago it was split from its ancestral pool: in other words, to deduce the current value of t given that we know N. Regardless of the properties of interest, we need to create a statistical framework within which to examine them. The statistical framework used is to conceptualise **genetic sampling** as multiple replicate populations stemming from the infinite pool and evolving according to the same rules. This is shown by animation in Figure 2.

In Figure 2:

- The infinite reference population contains five different allele types (different colours), with an equal proportion of each.
- We imagine the alleles in the infinite reference population mixing as if they were all put in a box and shaken.
- For a single replicate population of size N individuals, $2N$ alleles are drawn at random from the infinite pool to form **Generation 0**. These $2N$ alleles are numbered $1, 2, \ldots, 2N$ so that we can keep track of the descendents (or *lineages*) of each one.
- For later generations $t = 1, 2, \ldots$, we form generation t by sampling $2N$ alleles at random with replacement from generation $t - 1$.
- The whole process is repeated (conceptually) for many replicate populations.

The replicate populations form the key **statistical population** in our genetic models. Although it is often not made explicitly clear, when people talk about means and variances of genetic processes, they will typically be talking about means and variances with respect to the population of replicate populations. For example:

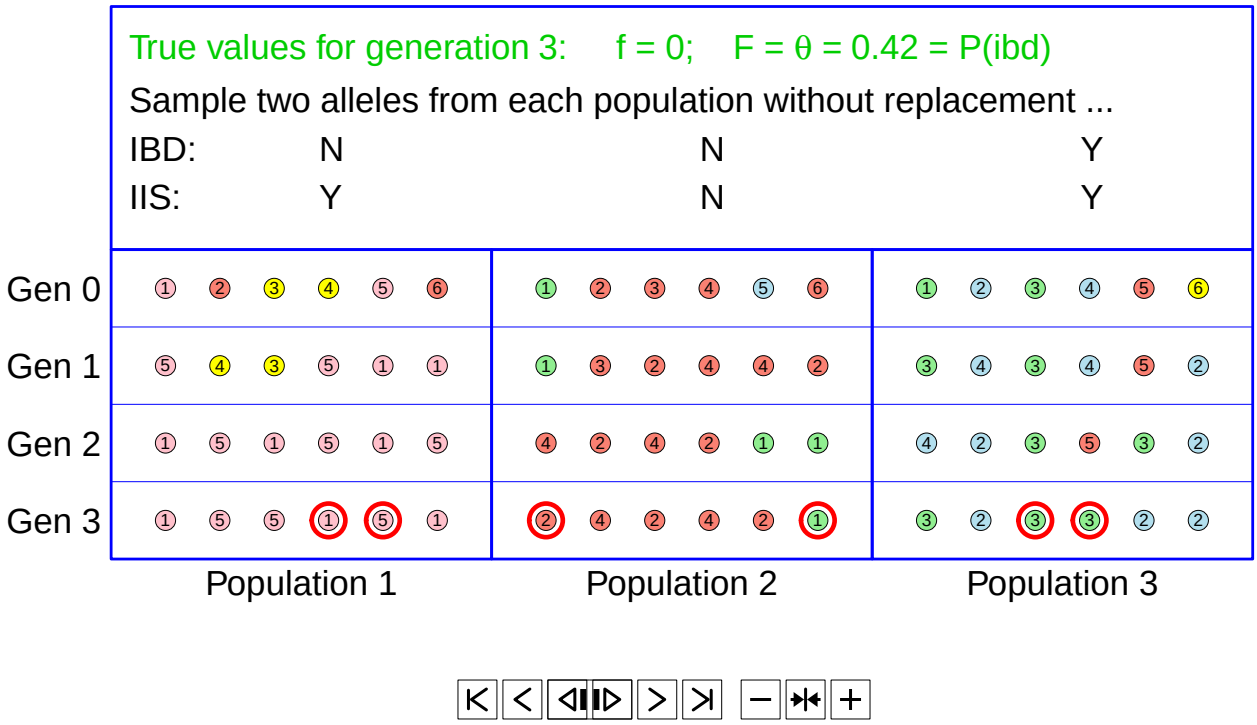


Figure 2: The Wright-Fisher model, showing evolution of multiple replicate populations of size N, each drawn at time 0 from the infinite reference population and allowed to evolve independently for generations 0, 1, 2, 3. Different colours correspond to different allele types, and numbers correspond to allele lineages within a population.

Statement:

'The expected frequency of allele A in generation t is $p_A = 1/5$.'

Meaning:

Imagine r replicate populations at time t. For each population $i = 1, 2, \ldots, r$, draw a single allele from generation t. Let X_i be the indicator for whether the selected allele is allele A: $X_i = 1$ if the allele from population i is allele A, and 0 otherwise. Then $\mathbb{E}(X_i) = p_A = 1/5$ for $i = 1, 2, \ldots, r$; and as $r \to \infty$, we have $\frac{1}{r}\sum_{i=1}^{r} X_i \to p_A = 1/5$.

The critical thing here is that the 'allele frequency' $p_A = 1/5$ is ***not*** referring to what is going on in any of the individual replicate populations. It is the allele frequency in the **infinite ancestral population**, which may never have existed, except in our heads.

Key points to notice from Figure 2 are:

- Small populations very quickly lose genetic diversity. Of the original five allele types, only one or two remain by generation 3, in each of the replicates shown.

- The population replicates may differ greatly in their genetic composition from each other and from the infinite reference population. In the Figure, the different replicates barely have any allele types in common.

Probability of identity by descent

In practice, we observe a small number of present-day replicate populations, from which we want to deduce properties of the overall genetic process from infinite reference to the present day: for example, for how many generations t the populations have been separated. The interest is likely to be in N and t, and allele frequencies in the reference population are just a nuisance. For this reason, a key idea in the study of genetic structure is that of ***identity by descent:***

Definition: Two alleles in a replicate population are **identical by descent (IBD)** if they stem from the same ancestor allele from generation 0.

In the animations, two alleles are identical by descent if they are marked with the same number: it means they are part of the same lineage.

Contrast this with ***identity in state***, which expresses whether two alleles have the same type or state:

Definition: Two alleles are **identical in state (IIS)** if they have the same allele type.

In the animations, two alleles are identical in state if they have the same colour.

The **probability of identity-by-descent** is the probability that two different alleles sampled from the same population at generation t are identical by descent. As before, the ***probability*** refers to the probability across all replicate populations.

Statement:

'The probability of identity by descent in generation 3 is 0.42.'

Meaning:

Imagine r replicate populations. For each population $i = 1, 2, \ldots, r$, draw two alleles, at random without replacement, from generation $t = 3$. Let X_i be the indicator for whether the two selected alleles are identical by descent: $X_i = 1$ if the alleles from population i are IBD, and 0 otherwise. Then $\mathbb{P}(\text{IBD}) = \mathbb{E}(X_i) = 0.42$ for $i = 1, 2, \ldots, r$; and as $r \to \infty$, we have $\frac{1}{r}\sum_{i=1}^{r} X_i \to \mathbb{P}(\text{IBD}) = 0.42$.

This concept of the probability of identity-by-descent is shown at the end of the animation in Figure 2.

The probability of identity-by-descent depends only on N and t in the Wright-Fisher model: it will be high when N is small, and it increases as t gets larger. It therefore encapsulates information of interest about the overall genetic process.

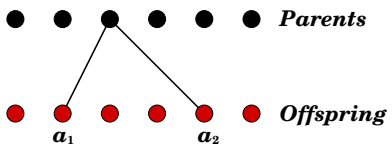
Result: For the Wright-Fisher model with N diploid individuals, the probability of identity-by-descent in generation t is:

$$\mathbb{P}(\text{IBD}) = 1 - \left(1 - \frac{1}{2N}\right)^t \quad \text{for } t = 0, 1, 2, \ldots$$

Proof: Let θ_t be the probability of identity-by-descent in generation t. Clearly, $\theta_0 = 0$ because all alleles start their own lineages in generation 0.

Consider generation t. We draw two alleles at random without replacement from the $2N$ available at time t. Call them a_1 and a_2.

The two parent alleles of a_1 and a_2 correspond to two random draws *with replacement* from the $2N$ alleles available at time $t-1$. With probability $\frac{1}{2N}$, alleles a_1 and a_2 have the same parent allele, so they are definitely IBD.



With probability $1 - \frac{1}{2N}$, alleles a_1 and a_2 have two *different* parents. In that case, a_1 and a_2 are IBD if and only if their parents are IBD. The probability that their parents are IBD is exactly the probability that two *different* alleles at generation $t-1$ are IBD, which is θ_{t-1}.

Thus θ_t, the probability that a_1 and a_2 are IBD, satisfies

$$\begin{aligned} \theta_0 &= 0 \\ \theta_t &= \frac{1}{2N} \times 1 + \left(1 - \frac{1}{2N}\right) \theta_{t-1} \qquad (t = 1, 2, 3, \ldots) \end{aligned}$$

Solving the difference equation gives the result:

$$\theta_t = 1 - \left(1 - \frac{1}{2N}\right)^t . \qquad \square$$

***Exercise* 1:** *(Answers on page 40.)* The expression $\theta_t = 1 - \left(1 - \frac{1}{2N}\right)^t$ can be derived directly without using the difference equation, by noting that $1 - \theta_t = \left(1 - \frac{1}{2N}\right)^t$. Find an argument to explain this expression.

Wright's F-statistics for genetic structure

Sewall Wright (1889-1988) was one of the founders of population genetics. He introduced a series of so-called ***F-statistics*** or ***Fixation Indices***. These quantities are very widely used today to summarise genetic structure. There are multiple different ways of defining and conceptualising these quantities, and the literature can be extremely confusing. Because they are such fundamental concepts of population genetics, we will try to draw together several common ways in which F-statistics are described.

F-statistics are defined with respect to the ***total population***, which is the infinite reference population; and the ***subpopulations***, which are individual replicate populations. The three F-statistics are:

- F_{IS} : inbreeding coefficient of individuals (I) within the subpopulation (S);
- F_{ST} : fixation index of subpopulations (S) within the total population (T);
- F_{IT} : overall inbreeding coefficient of individuals (I) within the total population (T).

However, Weir and Cockerham (1984) point out that there is confusion about whether F-statistics should indeed be regarded as *statistics* (functions of observations), or as *parameters*. It makes more sense to think of these measures of genetic structure as parameters, that we wish to estimate using data. To try to alleviate the confusion, they introduce new terminology:

- f : parameter corresponding to F_{IS} ;
  - within-population inbreeding coefficient;
  - correlation of genes within individuals within populations.

- θ : parameter corresponding to F_{ST} ;
  - ***coancestry coefficient***;
  - correlation of genes in different individuals in the same population;
  - probability of identity-by-descent for alleles selected from different individuals in the same population.

- F : parameter corresponding to F_{IT} ;
  - overall inbreeding coefficient;
  - correlation of genes within individuals.
  - probability of identity-by-descent for alleles selected from the same individual.

Weir and Cockerham's conventions have not been universally adopted, and much confusion still persists. We adopt the following system:

> $F = F_{IT}$, $\theta = F_{ST}$, and $f = F_{IS}$ are alternative names for the same ***parameters***.

The connections between F-statistics, probabilities of identity-by-descent, correlation of genes, inbreeding, and coancestry, will probably seem obscure. We explain these now.

The coancestry coefficient, θ, or F_{ST}

In the Wright-Fisher model, we did not organise the $2N$ alleles into N individuals. The coancestry coefficient, θ or F_{ST}, is the only one of the three parameters that does not involve individuals, so it is the only one that is directly applicable to the unmodified Wright-Fisher model. It also offers a good example for explaining why so many seemingly-different terms describe the same thing.

Definition: In the Wright-Fisher model, where there is no organisation of alleles into individuals, **$\theta = F_{ST}$ is the probability that two different alleles selected at random are identical by descent**, where the probability is understood to be taken across replicate populations as on page 6. At generation t, this gives (page 7):

$$\theta = F_{ST} = \mathbb{P}(\text{IBD}) = 1 - \left(1 - \frac{1}{2N}\right)^t .$$

For models that do organise alleles into individuals, **$\theta = F_{ST}$ is the probability that two alleles selected at random from *different* individuals are identical by descent.**

θ as correlation between genes

We cannot observe whether or not two alleles are identical-by-descent. We can determine allele *state* (denoted by colours on Figure 2), but not the ancestry of an allele. Even over one generation, by looking only at your own alleles, there is no way of telling which you got from your mother and which from your father. The only information we can deduce about the probabilities of identity-by-descent must come indirectly by examining allele states.

Consider any allele type A, whose proportion in the infinite reference population is p_A. Define

$$\mathcal{P}_{A/A} = \mathbb{P}(\text{two alleles selected at random from different individuals are both } A),$$

where the probability is taken across replicate populations at generation t, as usual. Thus we imagine drawing two alleles, a_1 and a_2, from two different individuals in a replicate population. $\mathcal{P}_{A/A}$ is the probability that a_1 and a_2 both have allele state A.

Recall that θ is the probability of identity-by-descent for the same two selected alleles a_1 and a_2. To link $\mathcal{P}_{A/A}$ and θ, consider the following two possibilities:

1. a_1 and a_2 are IBD (probability θ). If a_1 and a_2 are IBD, they definitely have the same state. The probability that this state is A is the probability that the original ancestor of a_1 and a_2 in generation 0 has state A, which is p_A. Note that p_A is the frequency of allele A in the *infinite reference population*, not in the replicate population from which a_1 and a_2 are drawn.

2. a_1 and a_2 are not IBD (probability $1 - \theta$). In that case, they have two different ancestors in generation 0, each of which were independently drawn from the infinite reference population. The probability that a_1 and a_2 are both of state A in this case is therefore p_A^2.

Putting these possibilities together gives:

$$\begin{aligned} \mathcal{P}_{A/A} &= \mathbb{P}(\text{two alleles drawn from different individuals are both of state } A) \\ &= \theta p_A + (1-\theta)p_A^2 \\ &= \theta p_A(1-p_A) + p_A^2\,; \\ \Rightarrow \quad \theta &= \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)}\,. \end{aligned} \tag{1}$$

This expression for θ holds for any allele state A, and for any genetic model (not just the Wright-Fisher model). Weir (1996) takes it as the definition of θ. An advantage of this expression is that we can at least begin to imagine using it to estimate θ, because it depends upon observable allele states rather than unobservable identity-by-descent. However, the frequency p_A refers to the *infinite reference population* — which quite possibly never existed — so we still have some work to do to make this expression useful.

From equation (1), we can quickly see that θ is the correlation of alleles of different individuals, in the following sense. For replicate populations $i = 1, 2, \ldots, r$, select two alleles from different individuals within each population at generation t. Let X_{i1} and X_{i2} be the indicators for whether the two alleles from population i have state A: $X_{ij} = 1$ if the jth allele has state A and 0 otherwise, for $j = 1, 2$. Then

$$\mathbb{E}(X_{i1}) = \mathbb{E}(X_{i2}) = \mathbb{P}(\text{ancestor allele has state } A) = p_A\,,$$

and

$$\mathbb{E}(X_{ij}^2) = \mathbb{E}(X_{ij}) = p_A\,;$$

so

$$\text{var}(X_{i1}) = \text{var}(X_{i2}) = p_A(1-p_A)\,.$$

Also,

$$\mathbb{E}(X_{i1}X_{i2}) = \mathbb{P}(\text{both drawn alleles have state } A) = \mathcal{P}_{A/A} \quad \text{by definition of } \mathcal{P}_{A/A}\,.$$

Thus,

$$\text{corr}(X_{i1}, X_{i2}) = \frac{\mathbb{E}(X_{i1}X_{i2}) - \mathbb{E}(X_{i1})\mathbb{E}(X_{i2})}{\sqrt{\text{var}(X_{i1})\text{var}(X_{i2})}} = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)} = \theta\ . \tag{2}$$

Summary so far...

We currently have:

$$\begin{aligned} \theta = F_{ST} &= \mathbb{P}(\text{IBD for alleles in different individuals in the same population}) \\ &= \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)} \\ &= \text{correlation of genes in different individuals in the same population.} \end{aligned}$$

These are true in all situations and for any allele type A. The same value of θ applies to all allele types.

For the Wright-Fisher model, we also have $\theta = 1 - \left(1 - \frac{1}{2N}\right)^t$.

R simulations

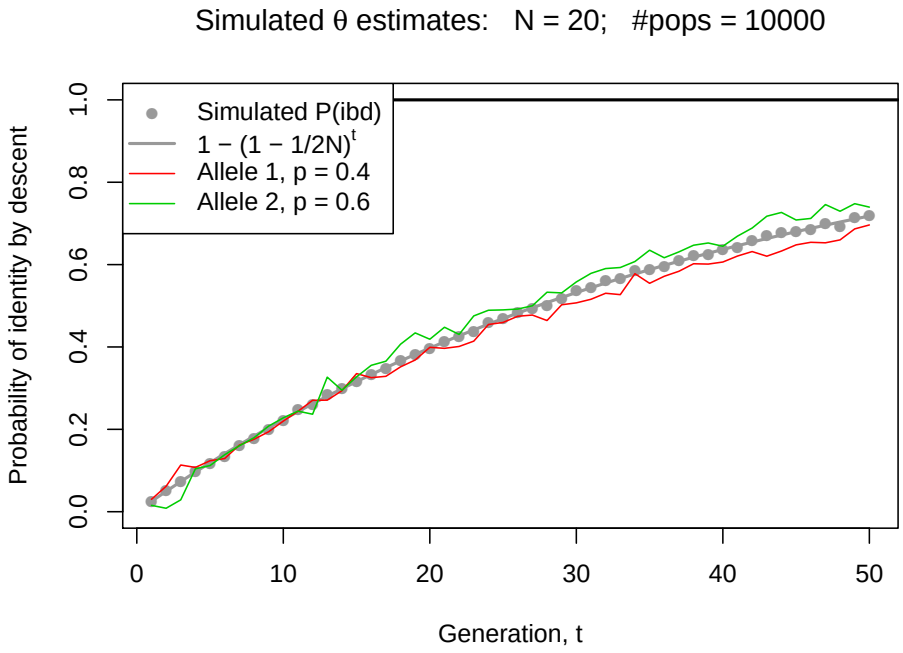


Figure 3: Simulations of the Wright-Fisher model.

Command: `wrightfisher.func(N=20, tmax=50, p.ancestral=c(0.4, 0.6), nrep=10000)`

This simulates 10,000 replicate populations of size $N = 20$ individuals under the Wright-Fisher model. There are two allele types, 1 and 2, with frequencies 0.4 and 0.6 in the infinite ancestral population. Two alleles are selected at random without replacement from the generation at time t ($t = 1, 2, \ldots, 50$) for each of the replicate populations, and the following quantities are logged:

- identity-by-descent: yes or no for each of the 10,000 populations.
- identity-in-state: yes or no for allele type 1, for each of the 10,000 populations; and yes or no for allele type 2, for each of the 10,000 populations.

The plot shows:

- Grey line: the curve $1 - \left(1 - \frac{1}{2N}\right)^t$ for $t \in \{0, 1, \ldots, 50\}$.
- Grey points: simulation estimates of $\mathbb{P}(\text{IBD})$: the number of the 10,000 populations for which the two selected alleles were IBD, divided by 10,000.
- Red line and green line: simulation estimates of $\frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)}$ for allele types 1 (red) and 2 (green). For allele type 1, $\mathcal{P}_{A/A}$ is the number of the 10,000 populations for which the two selected alleles were IIS for allele 1, divided by 10,000. Similarly for allele type 2.

We can see that each of the three simulation estimates follow the same curve, $1-\left(1-\frac{1}{2N}\right)^t$. The most accurate is $\mathbb{P}(\text{IBD})$ (why?) but this could not be used in practice as it is unobservable.

Practical Session 1

Explore the Wright-Fisher model in R by animation and simulation.

Animation:

```
wrightfisher.animate.func()
wrightfisher.animate.func(N=3, tmax=3, p.ancestral=c(0.4, 0.6),
       nrep=3, manual=F, skip.mix=F, delay.long=1, delay.short=0.5)
```

- `N` : number of individuals. Keep it small, as $2N$ alleles need to be displayed.
- `tmax` : number of generations simulated. Keep it small for display purposes.
- `p.ancestral` : allele frequencies p_A in the ancestral population. You can choose as many as you like. Values will be rounded to 2d.p. for the simulation. If the frequencies don't sum to 1, they will be rescaled.
- `nrep` : number of replicate populations to simulate. Keep it small for display purposes.
- `manual` : if True, animation will pause until you manually press Enter at several points.
- `skip.mix` : if True, the display of the initial mixing phase is omitted.
- `delay.long`, `delay.short` : delay in seconds between various phases of the animation.

For the fastest animation, use `skip.mix=T, delay.long=0, delay.short=0`.

Simulation: Produces graphics similar to Figure 3.

```
wrightfisher.func()
res = wrightfisher.func(N=50, tmax=50, p.ancestral=c(0.1, 0.2, 0.3, 0.4), nrep=1000)
```

The results in `res` have components:

- `$ibd` : simulation estimates of the probability of identity-by-descent: the grey points on the plot.
- `$PA.A` : simulation estimates of $\mathcal{P}_{A/A}$. A matrix with one row for each allele type A, and one column for each generation t.
- `$theta.by.allele.type` : simulation estimates of $\frac{\mathcal{P}_{A/A}-p_A^2}{p_A(1-p_A)}$. A matrix with one row for each allele type, and one column for each generation t. The matrix rows correspond to the coloured lines on the plot. For each t, the matrix columns are multiple estimates of the same quantity (θ_t), and the corresponding entry of `$ibd` is yet another estimate of θ_t.

***Exercise* 2:** *(Answers on page 40.)* Using the simulation `wrightfisher.func`, how would you change the arguments `N`, `tmax`, `p.ancestral`, and `nrep` to obtain the following results?

(a) Reduce $\theta = F_{ST}$ at generation $t = 50$?

(b) Increase $\theta = F_{ST}$ when $N = 20$?

(c) Increase the scatter of the grey points, $\mathbb{P}(\text{IBD})$, about the grey line $1 - \left(1 - \frac{1}{2N}\right)^t$?

(d) For a fixed `nrep=1000`, increase or decrease the deviation of the coloured lines $\frac{\mathcal{P}_{A/A}-p_A^2}{p_A(1-p_A)}$ from the grey line $1 - \left(1 - \frac{1}{2N}\right)^t$?

***Exercise* 3:** *(Answers on page 40.)* Consider an allele type A under the Wright-Fisher model for a population with $2N$ alleles. Define Y_t to be the number of alleles of type A in generation t. Clearly, $\{Y_t\}_{t\geq 0}$ is a Markov chain.

(a) Let $P = (p_{ij})$ be the transition matrix of the Markov chain. Find $p_{ij} = \mathbb{P}(Y_{t+1} = j \,|\, Y_t = i)$.

(b) Will the Markov chain converge to an equilibrium distribution as $t \to \infty$? If so, what are the equilibrium states?

(c) Show that $\{Y_t\}_{t\geq 0}$ is also a martingale; in other words that $\mathbb{E}(Y_{t+1} \,|\, Y_t, \ldots, Y_0) = Y_t$.

(d) When the chain reaches the state $Y_t = 0$ or $Y_t = 2N$, it is said to have reached *fixation*. Let T be the generation at which fixation occurs, and let α be the probability that the chain is eventually fixed at state $2N$ rather than state 0. Because $\{Y_t\}_{t\geq 0}$ is a bounded martingale, we can apply the Optional Stopping Theorem:

$$\mathbb{E}(Y_T) = \mathbb{E}(Y_0).$$

Using this, find α, the probability that all individuals eventually have allele A.

Population differentiation and fixation

It is clear that $\theta = F_{ST}$ is a measure of population differentiation or 'distance'. Under the Wright-Fisher model, all populations will eventually reach *fixation*: the allele frequencies will drift until there is only one allele type left in each population. The probability (across replicate populations) that a population is eventually fixed for allele A is p_A: see Exercise 3.

As $t \to \infty$, the probability of fixation approaches 1. F_{ST} or θ also approaches 1, which can be seen in two ways. Firstly,

$$\theta = 1 - \left(1 - \frac{1}{2N}\right)^t \to 1 \text{ as } t \to \infty\,.$$

Secondly, note that when the population is fixed at A with probability p_A, and fixed at some different allele with probability $1 - p_A$, then $\mathcal{P}_{A/A} = p_A$. This is because two alleles drawn from the population will both be A if and only if the population is fixed for A. Thus,

$$\theta = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)} \to \frac{p_A - p_A^2}{p_A(1-p_A)} = 1.$$

If the only process going on is genetic drift, θ tells us how long the populations have been separated:

$$t = \frac{\log(1-\theta)}{\log\left(1 - \frac{1}{2N}\right)}\,.$$

A more interesting situation is the *Infinite Islands Model with migration*, in which θ can be used to estimate the migration rate in a mainland-island situation.

Infinite Island Model: migration-drift equilibrium

Wright's Infinite Island Model is largely the same as the Wright-Fisher model, with one exception. When an allele is drawn for generation t of a replicate population, it is drawn in the usual way from the time $t-1$ alleles only with probability $1-m$. With probability m, the allele is a 'migrant': it is drawn directly from the infinite reference population.

This model is shown in the animation in Figure 4. Migrant alleles are marked with red squares, and are given a new number designating the start of a new lineage in the population. The model can be thought of as representing a large mainland population (the 'infinite reference'), with a large number of islands of equal population sizes. The islands are mostly isolated, but get occasional migrants from the mainland.

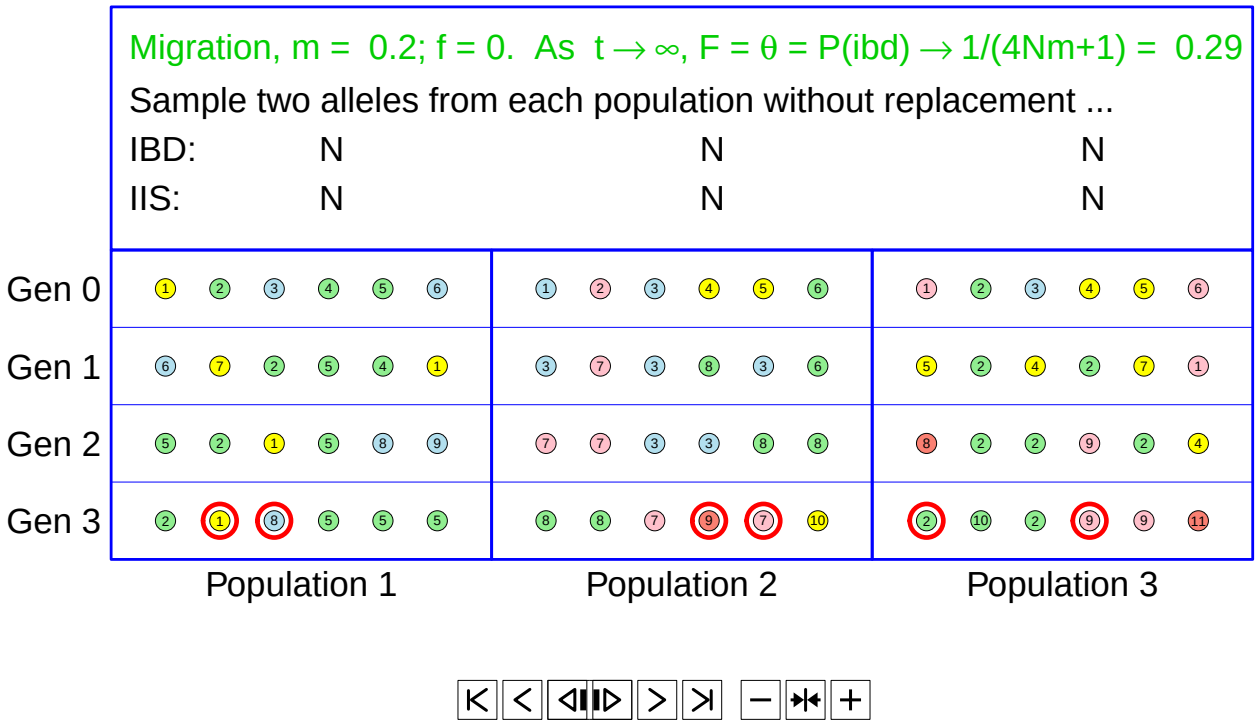


Figure 4: The Infinite Island model with migration parameter $m = 0.2$. Migrant alleles are marked during the animation with red squares, and are drawn directly from the infinite reference population.

The addition of migrants ensures that the populations will *not* reach fixation, and θ will *not* inevitably tend to 1. Instead, θ will reach an equilibrium value, representing a balance between the two opposing effects of (i) drift, which tends to increase θ, and (ii) migration, which tends to decrease θ. Effectively, the island populations do not drift so far apart from each other, because they are linked by a common connection with the mainland. The populations are said to reach ***migration-drift equilibrium.***

The infinite island model is particularly nice because there is a direct relationship between the equilibrium value of $\theta = F_{ST}$ and the migration rate, m. For this reason, θ is often used to estimate migration rate, m. You will often hear $\theta = F_{ST}$ referred to as a **measure of gene flow**. This refers to the following relationship.

Result: For the infinite island model with populations of size N individuals and migration rate m, the value of $\theta = F_{ST}$ at migration-drift equilibrium is approximately

$$\theta_\infty \simeq \frac{1}{4Nm+1}\,. \tag{3}$$

Proof: We use the definition of θ as the probability of identity by descent. Let θ_t be the probability of identity-by-descent in generation t. Draw two different alleles a_1 and a_2, from the $2N$ available at time t. Alleles a_1 and a_2 can only be IBD if they are both *non*-migrants: probability $(1-m)^2$. For two non-migrants, we draw from the $2N$ parents at time $t-1$ exactly as in the Wright-Fisher model on page 7. Thus, by partitioning into the cases of a_1 and a_2 sharing the same parent (probability $\frac{1}{2N}$) or not, and following the argument on page 7:

$$\mathbb{P}(a_1 \text{ and } a_2 \text{ are IBD} \,|\, \text{both non-migrants}) = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\theta_{t-1}\,.$$

Thus,

$$\theta_t = (1-m)^2\left\{\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\theta_{t-1}\right\}\,. \tag{4}$$

At migration-drift equilibrium, $\theta_t = \theta_{t-1} = \theta_\infty$. Rearranging (4):

$$\begin{aligned}\theta_\infty = \frac{(1-m)^2}{2N-(1-m)^2(2N-1)} &= \frac{1}{2N(1-m)^{-2} - 2N + 1} \qquad (5)\\ &= \frac{1}{2N\{1+2m+O(m^2)\} - 2N + 1} \quad \text{(Taylor expansion)}\\ &\simeq \frac{1}{4Nm+1} \quad \text{as required.}\end{aligned}$$

Note that for large m, the exact expression (5) can be used. □

All models are wrong, but some are useful...

The infinite island model is very clearly an oversimplification of reality. There is some conflict of opinion about how useful equation (3) is for estimating gene flow. No independent opinion is attempted here, but the following quote from Neigel (2002) is relevant:

No one has seriously argued that natural populations have [the infinite island model] characteristics; the model is just a convenient abstraction that isolates the opposing effects of genetic drift and gene flow. However the model is relevant to the interpretation of data from real populations because it is possible to relax its assumptions without greatly altering the relationship

between F_{ST} and Nm. The number of populations doesn't really have to be infinite (or even a very large number), mutation and selection are only likely to be important when populations are very large, and even if gene flow is limited by distance the overall value of F_{ST} is expected to be similar to that obtained for the infinite island model. ... The infinite island model has provided a robust, albeit coarse guide to how overall levels of gene flow influence overall F_{ST}.

In practice, using the expression $\theta = \frac{1}{4Nm+1}$ to estimate migration rate, m, requires an estimate of the population size N (or rather, *genetic effective population size*, which is even more difficult to estimate). Instead, the expression is often used to estimate the composite Nm, which corresponds to the (effective) number of migrants per population per generation.

R simulations

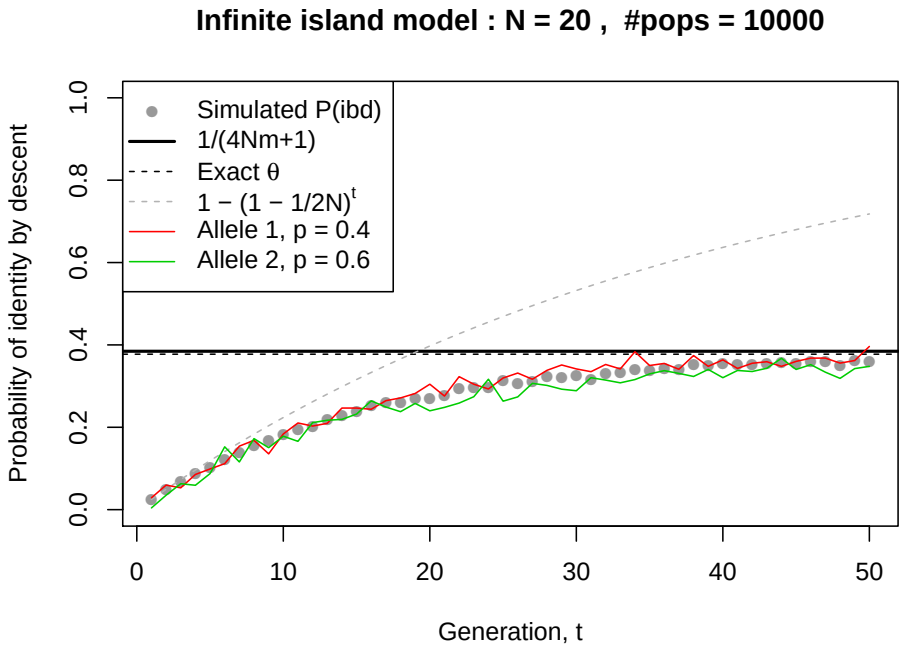


Figure 5: Simulations of the Infinite Island model with $N = 20$, $m = 0.02$. The bold horizontal line shows the approximate equilibrium value $\theta = 1/(4Nm+1)$, with the exact expression given by (5) shown by the dashed horizontal line. The dashed grey line shows the Wright-Fisher value of θ without migration for comparison.

Command:
`infinite.island.func(N=20, m=0.02, tmax=50, p.ancestral=c(0.4,0.6), nrep=10000)`
For this population size $N = 20$, a small amount of migration is sufficient to lower the equilibrium θ from 1 to $1/(4Nm + 1) = 0.38$. The migration rate is $m = 0.02$, corresponding to an average of less than one migrant allele per generation in each population.

Inbreeding Model

So far in the Wright-Fisher model and Infinite Island model, we have not organised the $2N$ alleles into N individuals. We have focused on the coancestry coefficient, $\theta = F_{ST}$, and not described the other two of Wright's F-statistics, $f = F_{IS}$ and $F = F_{IT}$. The **inbreeding model** is more general than the Wright-Fisher model, as it does organise alleles into individuals, and allows the two alleles of a single individual to be more closely related than two alleles of two different individuals. This effect is called **inbreeding**:

Definition: A population exhibits **inbreeding** if the two alleles of a single individual are more likely to be the same than two alleles selected from two different individuals in the population.

This suggests a connection with the idea of correlation of alleles within individuals, similar to the descriptions of θ as correlation of alleles of different individuals seen on page 9.

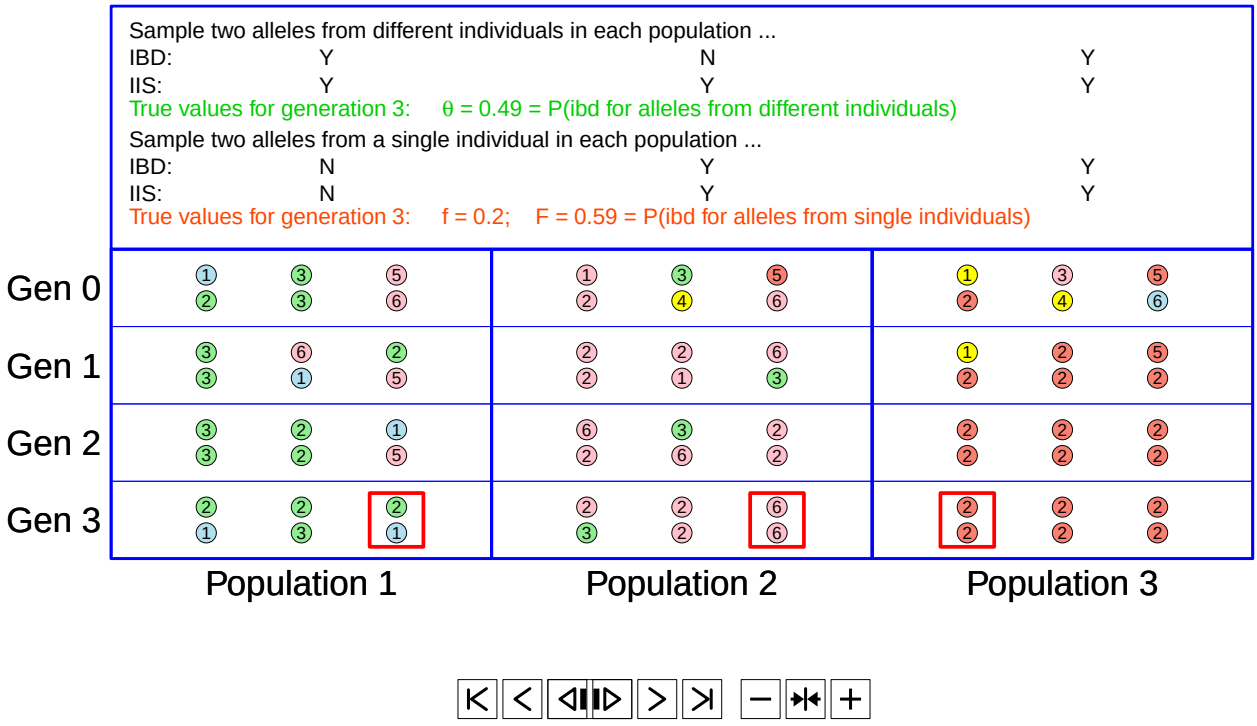


Figure 6: The inbreeding model with inbreeding parameter $f = 0.2$. 'Inbred' individuals are marked during the animation with red rims, and always have two identical alleles.

The inbreeding model is similar to the Wright-Fisher model, but differs as follows.

- Organise the $2N$ alleles in each population into N individuals.
- When drawing the two alleles for an individual in generation t, follow the usual Wright-Fisher scheme with probability $1 - f$, by drawing the two alleles at random with replacement from those available at time $t - 1$.
- With probability f, the individual is *'inbred'*. In this case, draw only one allele from time $t - 1$ for this individual, and give it two copies of the same allele.
- f is called the ***within-population inbreeding coefficient***, and is Wright's F_{IS}.

As ever, this mechanism for conceptualising inbreeding is not thought to be realistic, but it is a plausible description of real genetic structure.

Wright's F-statistics revisited

Before looking at the inbreeding model in more detail, we revisit Wright's F-statistics and focus on the two statistics relating to inbreeding: F_{IT} and F_{IS}. The definitions below are valid for any genetic model, not just the inbreeding model which we will return to soon. However, as always, they do rely on the idea of an infinite reference population and *genetic sampling* leading to a population of replicate populations. It is the statistical properties of the replicate populations that we are interested in.

Overall inbreeding coefficient, $F = F_{IT}$

The concept and derivations of the overall inbreeding coefficient, F or F_{IT}, are exactly analogous to those for $\theta = F_{ST}$, except that F relates to alleles within the same individual, whereas θ relates to alleles of different individuals within the same population.

Probability of identity by descent

- $\theta = F_{ST}$ is the probability of identity-by-descent for two alleles selected from *different* individuals in the same population: page 9. The probability is taken across all replicate populations.

- $F = F_{IT}$ is the probability of identity-by-descent for two alleles selected from *a single* individual in a population. Again, the probability is taken across all replicate populations.

  In the Wright-Fisher model or infinite island model, alleles are not organised into individuals. For these two models, therefore, there is no difference between selecting two alleles from the same individual or from different individuals, so $F = \theta$. In the inbreeding model, $F \neq \theta$.

Relationship to probabilities of identity in state

- $\theta = F_{ST} = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)}$, where $\mathcal{P}_{A/A}$ is the probability (across populations) that two alleles selected from *different* individuals in the same population are both of allele type A. The expression is valid for any allele type A, regardless of its frequency p_A in the ancestral population.

  Recall (page 9): selected alleles a_1 and a_2 are definitely identical in state if they are IBD: probability θ; so with probability θp_A, they are IBD and have allele type A. If they are not IBD, they have two different ancestors at generation 0, so the probability they are IIS for allele A is p_A^2. So

$$\mathcal{P}_{A/A} = \theta p_A + (1-\theta)p_A^2 \qquad \Rightarrow \qquad \theta = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)}\,.$$

- The analogous expression for F is

$$F = F_{IT} = \frac{\mathcal{P}_{AA} - p_A^2}{p_A(1-p_A)}\,,$$

valid for any allele type A, where $\mathcal{P}_{AA}$ is the probability (across populations) that both alleles of a *single* individual are of type A.

Explanation: Let a_1 and a_2 be the two alleles of the selected individual. By definition of F, the probability that a_1 and a_2 are IBD is F. The probability that a_1 and a_2 are IBD and have allele type A is therefore Fp_A, because p_A is the probability that their common ancestor allele in generation 0 had type A.

With probability $1-F$, the alleles a_1 and a_2 are not IBD. They therefore had two different ancestor alleles at time 0, drawn independently from the reference population. Thus the probability they are both of type A is p_A^2.

Overall,

$$\mathcal{P}_{AA} = Fp_A + (1-F)p_A^2 \qquad \Rightarrow \qquad F = \frac{\mathcal{P}_{AA} - p_A^2}{p_A(1-p_A)}\,.$$

Correlation of alleles within or between individuals

- $\theta = F_{ST}$ is the correlation between two alleles selected from *different* individuals in the same population.

Recall (page 9): let X_1 and X_2 be indicator random variables for whether alleles a_1 and a_2 have type A; where a_1 and a_2 are selected from different individuals in the same population. Then

$$\mathbb{E}(X_i) = \mathbb{E}(X_i^2) = p_A \;\; (i=1,2); \quad \mathbb{E}(X_1X_2) = \mathcal{P}_{A/A} \text{ by definition of } \mathcal{P}_{A/A}\,.$$

So

$$\text{corr}(X_1, X_2) = \frac{\mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)}{\mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2} = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)} = \theta.$$

- Similarly, let Y_1 and Y_2 be indicator random variables for whether alleles a_1 and a_2 belonging to a *single* individual have type A. Then

$$\mathbb{E}(Y_i) = \mathbb{E}(Y_i^2) = p_A \;\; (i=1,2); \quad \mathbb{E}(Y_1Y_2) = \mathcal{P}_{AA} \text{ by definition of } \mathcal{P}_{AA}\,.$$

So

$$\text{corr}(Y_1, Y_2) = \frac{\mathbb{E}(Y_1Y_2) - \mathbb{E}(Y_1)\mathbb{E}(Y_2)}{\mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2} = \frac{\mathcal{P}_{AA} - p_A^2}{p_A(1-p_A)} = F\,.$$

Overall, $\theta = F_{ST}$ is the correlation between alleles of different individuals in the same population; whereas $F = F_{IT}$ is the correlation of alleles within individuals. In either case, the correlation is taken across all replicate populations.

Within-population inbreeding coefficient, $f = F_{IS}$

To understand the within-population inbreeding coefficient, f, we need a new way of thinking. Up till now we have been thinking in terms of the *total population:* the 'population of populations' generated by genetic sampling. By contrast, the within-population inbreeding coefficient f refers to processes *inside* a replicate population, but it is still a global parameter that applies to all replicate populations at once.

The key difference that this makes is that we need to consider allele frequencies *within* populations, instead of considering the overall allele frequency p_A. Each generation, the allele frequency in replicate population i will change due to random genetic drift; so we also have to be clear which generation of allele frequency we are referring to.

Define the following quantities for a single replicate population i and for allele types A and B:

$$\begin{aligned} p_{Ai} &= \text{frequency of allele type } A \text{ in replicate population } i \text{ at generation } t \\ p_{Ai\star} &= \text{frequency of allele } A \text{ in replicate population } i \text{ in the } \textit{parental} \text{ generation } t-1. \\ \mathcal{P}_{AAi} &= \text{probability an individual in replicate population } i \text{ at generation } t \text{ has genotype } AA. \end{aligned}$$

Similarly, for another allele type B, define p_{Bi} and $p_{Bi\star}$ as the frequency of allele type B respectively in the offspring generation t and the parental generation $t-1$. Here, the allele frequency is the proportion of the $2N$ alleles with the corresponding type.

$f = F_{IS}$ as the probability of inbreeding

As in the description in the inbreeding model, we define f to be the probability that an individual's two alleles are identical copies of the same allele from the parental generation. Consider the genotype probabilities for individuals in generation t of population i.

With probability f, an individual has two copies of the same parental allele. The probability this allele is of type A is $p_{Ai\star}$.

With probability $1-f$, the individual's alleles are drawn at random with replacement from the parental stock. The probability that they are both of type A is $p_{Ai\star}^2$.

Thus the probability that the individual has genotype AA is

$$\mathcal{P}_{AAi} = f p_{Ai\star} + (1-f) p_{Ai\star}^2 \,.$$

For any two allele types A and B, the individual can only have genotype AB if it is not 'inbred'. Thus

$$\mathcal{P}_{ABi} = 2(1-f) p_{Ai\star} p_{Bi\star} \,.$$

The '2' arises because the selected alleles could be A first and B second, or the other way around.

These expressions hold for all allele types. If there are only two allele types available, A and B, then the entire set of genotype probabilities is:

$$\begin{aligned}\mathcal{P}_{AAi} &= f p_{Ai\star} + (1-f)\, p_{Ai\star}^2 \\ \mathcal{P}_{ABi} &= (1-f)\, 2\, p_{Ai\star}\, p_{Bi\star} \\ \mathcal{P}_{BBi} &= f p_{Bi\star} + (1-f)\, p_{Bi\star}^2\end{aligned}$$

(Check that these probabilities sum to 1 if $p_{Ai\star} + p_{Bi\star} = 1$.)

This shows that f describes the deviation from Hardy-Weinberg proportions in each generation. We can imagine that proportion $(1-f)$ of the population follows Hardy-Weinberg proportions, while proportion f is 'inbred'. The parameter f controls the ***surplus of homozygotes*** (individuals with both their alleles the same) compared with Hardy-Weinberg proportions.

$f = F_{IS}$ as the correlation of alleles within individuals within populations

Inside replicate population i, we consider selecting two alleles a_1 and a_2 from a single individual at generation t. Here, we have to be careful, because there are two things we might mean by 'selecting an individual at generation t':

1. Generation t is already formed, and we sample from the realised N individuals in generation t;
2. Generation t is *not yet formed* from the parental generation $t-1$. Instead, our sampling at generation t describes the process of *forming individuals* for generation t, and therefore the relevant frequencies are those in the parental generation, $p_{Ai\star}$.

Conventionally, it is Option 2 that is used. The sampling process therefore describes the ***uniting of gametes to form generation t***, rather than a sampling of the specific N individuals that survived from the infinitely many united gametes. For this reason, the correlations we are about to describe are often referred to as the ***correlation between uniting gametes***. This is a rather subtle way of saying that the correlations are calculated with respect to the allele frequencies in the parental generation, rather than the offspring generation.

Thus, let a_1 and a_2 be a pair of uniting gametes for generation t, in other words, the two alleles of a 'potential' individual at generation t. Let Y_1 and Y_2 be indicator random variables for whether alleles a_1 and a_2 have type A. Then

$$\mathbb{E}(Y_k \,|\, p_{Ai\star}) = \mathbb{E}(Y_k^2 \,|\, p_{Ai\star}) = p_{Ai\star} \;\; (k=1,2); \quad \mathbb{E}(Y_1 Y_2 \,|\, p_{Ai\star}) = f p_{Ai\star} + (1-f) p_{Ai\star}^2.$$

So

$$\operatorname{corr}(Y_1, Y_2 \,|\, p_{Ai\star}) = \frac{\mathbb{E}(Y_1 Y_2 \,|\, p_{Ai\star}) - \mathbb{E}(Y_1 \,|\, p_{Ai\star})\mathbb{E}(Y_2 \,|\, p_{Ai\star})}{\mathbb{E}(Y_k^2 \,|\, p_{Ai\star}) - \mathbb{E}(Y_k \,|\, p_{Ai\star})^2} = \frac{f p_{Ai\star} + (1-f) p_{Ai\star}^2 - p_{Ai\star}^2}{p_{Ai\star}(1-p_{Ai\star})} = f\,.$$

In this sense, the within-population inbreeding coefficient f is the correlation between uniting gametes at any generation.

The relationship between f, F, and θ

Result: Wright's three F-statistics, $f = F_{IS}$, $F = F_{IT}$, and $\theta = F_{ST}$, are linked by the following formula:

$$f = \frac{F - \theta}{1 - \theta}\,. \tag{6}$$

(This expression is stated without explanation on page 49 of Weir (1996).)

Proof: To link the global quantities F and θ with the within-population quantity f, we need to relate the global $\mathcal{P}_{AA}$ to its within-population analogue $\mathcal{P}_{AAi}$, by taking the expectation across populations.

Let $\mathbb{E}_\star$ denote the expectation over the distribution of $p_{Ai\star}$, the allele frequency in the parental distribution. First note the following:

- $\mathcal{P}_{AA} = \mathbb{E}_\star(\mathcal{P}_{AAi})$ by definition of $\mathcal{P}_{AA}$ as the probability across populations of sampling a homozygote for allele type A.
- $\mathbb{E}_\star(p_{Ai\star}) = p_A$.
- $\mathbb{E}_\star(p_{Ai\star}^2) = \mathcal{P}_{A/A}$. This is because $\mathcal{P}_{A/A}$ describes the probability of sampling two A alleles from two *different* individuals at generation t, so the two parental alleles at generation $t-1$ are sampled independently (with replacement) from the parental allele frequencies $p_{Ai\star}$.

Using these expressions, we have:

$$\begin{aligned}
\mathcal{P}_{AA} &= \mathbb{E}_\star(\mathcal{P}_{AAi}) \\
&= \mathbb{E}_\star\left\{f p_{Ai\star} + (1-f)p_{Ai\star}^2\right\} \quad \text{from above;} \\
&= f p_A + (1-f)\mathcal{P}_{A/A} \\
\Rightarrow \quad f &= \frac{\mathcal{P}_{AA} - \mathcal{P}_{A/A}}{p_A - \mathcal{P}_{A/A}} \\
\therefore \quad f &= \frac{F-\theta}{1-\theta} \qquad \text{using } F = \tfrac{\mathcal{P}_{AA}-p_A^2}{p_A(1-p_A)},\ \theta = \tfrac{\mathcal{P}_{A/A}-p_A^2}{p_A(1-p_A)}.
\end{aligned}$$

□

Other ways of defining f, θ, and F

As we've already seen, there are lots of different ways of defining and conceptualising the F-statistics $f = F_{IS}$, $\theta = F_{ST}$, and $F = F_{IT}$. In fact, there are even more ways in common use. For completeness, we describe these here.

f, θ, and F as measures of heterozygosity

This is perhaps the most common way of describing f, θ, and F, at least in biological texts. ***Heterozygosity*** describes the level of ***heterozygotes*** in a population. An individual is a heterozygote at a particular locus if its two alleles are of different types. If its two alleles are of the same type, it is a ***homozygote***. We have already seen that inbreeding inflates the homozygosity in a population. Even if there is no inbreeding ($f = 0$), a large θ implies that populations are reaching fixation, so the individuals inside the replicate populations are more likely to be homozygous than they would be within the infinite reference population, or *total* population. These ideas motivate the definitions of f, F, and θ in terms of heterozygosity levels.

The previous developments in terms of probabilities of identity-by-descent and correlation of alleles are statistically more precise, and biologically more general, as heterozygosity only applies to diploid organisms whereas the F-statistics (in particular $\theta = F_{ST}$) are relevant to any organisms. Even for diploid organisms, there is no concept of heterozygosity for their mtDNA haplotypes. However, heterozygosity is a fundamental biological concept because it affects population fitness: heterozygous individuals may have two chances of resistance to a disease, or some other useful trait. This is probably why it is often taken as the definition of choice for F-statistics.

Suppose at first that there are just two allele types, A and B. Define the following measures for allele A:

- $H_T = 2p_A(1 - p_A)$: expected heterozygosity in the *total* population, assuming random mating. The total population refers to the infinite reference population. Thus H_T is the frequency of genotype AB (heterozygotes) in the total population, if genotypes are formed by independent selection of alleles.
- $H_S = \mathbb{E}_\star \{2p_{Ai\star}(1 - p_{Ai\star})\}$: the mean, over replicate populations, of the expected heterozygosity within the populations, if there were random mating within the populations. As in the previous section, we consider the heterozygosity in *uniting gametes* for generation t, so we use the allele frequency $p_{Ai\star}$ from the parental generation $t - 1$ for replicate population i.

  From the results on page 22, we have $H_S = \mathbb{E}_\star \{2p_{Ai\star}(1 - p_{Ai\star})\} = 2(p_A - \mathcal{P}_{A/A})$.
- $H_I = \mathbb{E}_\star(\mathcal{P}_{ABi}) = \mathcal{P}_{AB}$, the mean *observed* heterozygosity within replicate populations. This allows for the possibility that mating is not random within replicate populations.

Now in the total population, we have

$$p_A = \mathcal{P}_{AA} + \frac{1}{2}\mathcal{P}_{AB}\,,$$

because we can imagine selecting an allele by first selecting an individual, and then selecting one of the individual's two alleles. If the individual has genotype AA (probability $\mathcal{P}_{AA}$), then the selected allele will definitely be of type A. If the individual has genotype AB (probability $\mathcal{P}_{AB}$), the selected allele will be of type A with probability $\frac{1}{2}$. Thus

$$H_I = \mathcal{P}_{AB} = 2(p_A - \mathcal{P}_{AA})\,.$$

Wright's F-statistics are then defined as follows:

$$f = F_{IS} = \frac{H_S - H_I}{H_S}\,, \qquad \theta = F_{ST} = \frac{H_T - H_S}{H_T}\,, \qquad F = F_{IT} = \frac{H_T - H_I}{H_T}\,. \tag{7}$$

With these expressions, it is easy to derive the relationship $f = \dfrac{F-\theta}{1-\theta}$.

***Exercise* 4:** Check that the expressions in (7) satisfy the previous definitions $\theta = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)}$, $F = \frac{\mathcal{P}_{AA} - p_A^2}{p_A(1-p_A)}$, $f = \frac{F-\theta}{1-\theta}$.

When there are more than two allele types, the expressions above still apply for each allele type A, where all other allele types are grouped into a single 'not-A' type.

$\theta = F_{ST}$ as the between-population percentage of variance

Another common definition of $\theta = F_{ST}$ is as the between-population component of variance. By now, you will be alert to the question 'variance of what?' The intended variance is the variance in allele selection, for the gametes forming generation t. Explicitly:

- Select one allele at random from any replicate population, from the gametes that will form generation t. Let X be the indicator for whether this allele is of type A.
- $\mathbb{E}(X) = p_A$, and $\mathbb{E}(X^2) = p_A$, so $\mathrm{var}(X) = p_A(1-p_A)$. This is the total variance of X.
- Partition the total variance of X into a within-populations component, and a between-populations component, using the law of total variance:

$$\mathrm{var}(X) = \mathbb{E}_\star\left\{\mathrm{var}(X \,|\, p_{Ai\star})\right\} + \mathrm{var}_\star\left\{\mathbb{E}(X \,|\, p_{Ai\star})\right\}\,.$$

  As previously, the expectation and variance are taken over the distribution of $p_{Ai\star}$, the allele frequency of allele type A in the parental generation $t-1$ of replicate population i.
- The between-populations component of the variance is the second term,

$$\mathrm{var}_\star\left\{\mathbb{E}(X \,|\, p_{Ai\star})\right\} = \mathrm{var}_\star(p_{Ai\star}) = \mathbb{E}_\star(p_{Ai\star}^2) - \left\{\mathbb{E}_\star(p_{Ai\star})\right\}^2 = \mathcal{P}_{A/A} - p_A^2\,,$$

  using $\mathbb{E}_\star(p_{Ai\star}^2) = \mathcal{P}_{A/A}$ as on page 22.

Viewing θ as the percentage of variance that is between populations implies taking the ratio of between-populations variance to total variance:

$$\theta = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)} \qquad \text{as before.}$$

Inbreeding model revisited

For the Wright-Fisher model, we had $f = 0$, so $F_t = \theta_t$ for any generation t, and $\theta_t = 1 - \left(1 - \frac{1}{2N}\right)^t$. The inbreeding model is a bit like the Wright-Fisher model but with inbreeding included. We can derive analogous expressions for θ_t and F_t.

Expressions for θ_t and F_t

Result: For the inbreeding model with size N individuals and within-population inbreeding parameter f, the generation-t values of θ and F are

$$\theta_t = 1 - \left\{1 - \frac{(1+f)}{2N}\right\}^t \qquad F_t = 1 - (1-f)\left\{1 - \frac{(1+f)}{2N}\right\}^t .$$

Proof: We set up coupled recursions for θ_t and F_t. For generation 0,

$$\theta_0 = 0 \qquad F_0 = f\,,$$

using the definitions as probabilities of identity-by-descent between and within individuals.

Consider θ_t, the probability two alleles a_1 and a_2 in different individuals at time t are IBD. Partition over different possibilities for the parent alleles of a_1 and a_2 at time $t-1$:

Parent alleles	Probability	$\mathbb{P}$(IBD) given these parents
Same parent allele	$\frac{1}{2N}$	1
Different parent alleles within same individual	$\frac{1}{2N}$	F_{t-1}
Different parent alleles in different individuals	$1 - \frac{2}{2N}$	θ_{t-1}

Thus

$$\theta_t = \frac{1}{2N} + \frac{1}{2N}F_{t-1} + \left(1 - \frac{1}{N}\right)\theta_{t-1}\,. \tag{8}$$

Similarly, consider F_t, the probability two alleles a_1 and a_2 in a single individual at time t are IBD. The individual can be 'inbred' (probability f), or with probability $1-f$ there are the same possibilities for the parents as before. Partition over possibilities:

Possibility	Probability	$\mathbb{P}$(IBD) given possibility
Inbred	f	1
Not inbred; same parent allele	$\frac{1-f}{2N}$	1
Not inbred; different parents in same individual	$\frac{1-f}{2N}$	F_{t-1}
Not inbred; different parents in different individuals	$(1-f)\left(1 - \frac{2}{2N}\right)$	θ_{t-1}

Thus

$$\begin{aligned} F_t &= f + (1-f)\left\{\frac{1}{2N} + \frac{1}{2N}F_{t-1} + \left(1 - \frac{1}{N}\right)\theta_{t-1}\right\} \\ &= \frac{1}{2N} + f\left(1 - \frac{1}{2N}\right) + \frac{(1-f)}{2N}F_{t-1} + (1-f)\left(1 - \frac{1}{N}\right)\theta_{t-1}\,. \end{aligned} \tag{9}$$

Equation (8) gives F_t in terms of $\{\theta_t\}$. Substituting for F_t and F_{t-1} in (9) gives a difference equation in θ_t. Solving gives the results. □

R simulations: inbreeding model

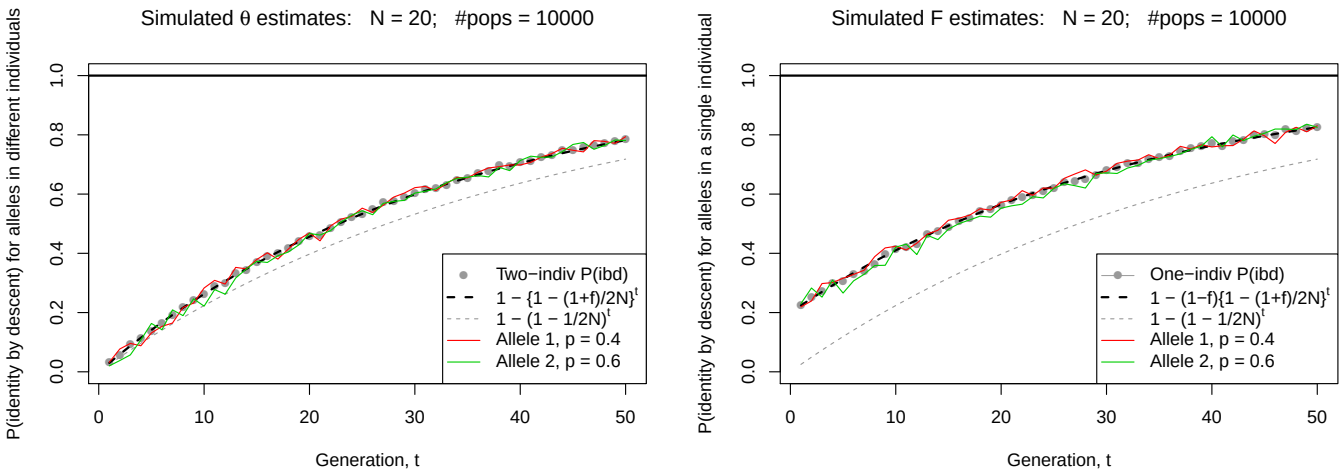


Figure 7: Simulations of the inbreeding model with $N = 20$, $f = 0.2$.

Command:
`inbreeding.model.func(N=20, f=0.2, tmax=50, p.ancestral=c(0.4, 0.6), nrep=10000)`
For this population size $N = 20$, the within-population inbreeding coefficient is quite large at $f = 0.2$. The effect is a small increase in θ over its Wright-Fisher value (shown in the dashed grey line), and a larger increase in F over its Wright-Fisher value, in the generations before convergence to 1. In the Wright-Fisher model without inbreeding, $F = \theta$.

Practical Session 2

1. Infinite Island model

Animations: Use `wrightfisher.animate.func` but with the parameter m added, where m is the probability that each individual is a migrant.

```
wrightfisher.animate.func(m=0.2)
## For faster simulation:
wrightfisher.animate.func(N=3, m=0.2, tmax=3, p.ancestral=c(0.4, 0.6),
      nrep=3, skip.mix=T, delay.long=0.2, delay.short=0.2)
```

Simulations: Produces graphics similar to Figure 5.

```
infinite.island.func()
infinite.island.func(N=50, m=0.1, tmax=50, p.ancestral=c(0.4, 0.6), nrep=1000)
```

***Exercise* 5:** *(Answers on page 41.)*

(a) Consider θ in the infinite island model with particular values of `m`, `N`, `p.ancestral`, and `nrep`. Which argument would you change, and how, to achieve the same value of θ in the Wright-Fisher model?

(b) Try the following simulations:

```
infinite.island.func(m=0.1, tmax=50, p.ancestral=c(0.4, 0.6))    # High m, N=50
infinite.island.func(m=0.01, tmax=50, p.ancestral=c(0.4, 0.6))   # Low m, N=50
infinite.island.func(m=0.01, tmax=100, p.ancestral=c(0.4, 0.6))  # Low m, longer t
```

Does θ_t converge more quickly to equilibrium with large m or with small m? Can you think of an argument to explain why? (Hint: consider the relationship between θ under the infinite island model, and θ under the Wright-Fisher model.)

(c) Run the animation function with $m = 1$: `wrightfisher.animate.func(m=1)`. What should the value of θ be? Why does the expression $1/(4Nm + 1)$ give the wrong answer?

2. Inbreeding model

Animations: e.g. `inbreeding.animate.func(N=3, f=0.2, tmax=3)`

Simulations: Produces graphics similar to Figure 7. Example:
`inbreeding.model.func(N=50, f=0.1, tmax=50, p.ancestral=c(0.4, 0.6), nrep=1000)`

***Exercise* 6:** *(Answers on page 42.)*

(a) If $f = 0$, what is the relationship between F and θ? What is the relationship between the inbreeding model and the Wright-Fisher model when $f = 0$?

(b) If we had data from only one replicate population, which quantities out of F, θ, and f would we have information to estimate?

(c) Suppose we have $f = 0$ in the inbreeding model. Sample a single individual from one population at time $t > 0$. Is this individual more likely to be a homozygote than it would be if it were sampled directly from the infinite reference population? Why?

(d) Run the following command (it will take a minute):
`res=inbreeding.model.func(N=10,f=0.6,tmax=10,p.ancestral=c(0.4,0.6),nrep=10000)`
The result `res` has components that include `pA`= $\mathbb{E}_{\text{pops, } i}(p_{Ai})$, `pAsq`= $\mathbb{E}_i(p_{Ai}^2)$, `PA.A`= $\mathcal{P}_{A/A} = \mathbb{E}_i(\mathcal{P}_{A/A,i})$, `PAA`= $\mathcal{P}_{AA} = \mathbb{E}_i(\mathcal{P}_{AAi})$, and `two.indiv.ibd`. Using plots, verify that $\mathbb{E}_\star(p_{Ai\star}^2) = \mathcal{P}_{A/A}$, where the $\star$ refers to the parental generation relative to $\mathcal{P}_{A/A}$. Also find the correct objects to verify that θ is the percentage of variance between populations. Would you expect that $\mathbb{E}_i(p_{Ai}^2) = \mathcal{P}_{AA}$?

Estimating $f = F_{IS}, \theta = F_{ST}$, and $F = F_{IT}$

It's now time to consider how we would use real data to estimate the three components of genetic structure, f, θ, and F. Although many estimating approaches have been proposed, it is probably true that the approach of Weir and Cockerham (1984) has the widest acceptance. Their approach is used (with a slight modification for combining estimates across loci) by the popular Genepop software (Raymond and Rousset, 1995; Rousset, 2008): see `http://genepop.curtin.edu.au/Appendix2.htm`.

Consider the definitions:

$$\theta = \frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)}\,, \quad F = \frac{\mathcal{P}_{AA} - p_A^2}{p_A(1-p_A)}\,, \quad f = \frac{F-\theta}{1-\theta}\,.$$

These expressions hold for any allele A. They rely upon p_A, the frequency of allele A in the ancestral population — which probably never existed and about which we have no data. Weir and Cockerham's approach is to derive quantities with expectations that depend upon F, θ, and $p_A(1-p_A)$, and to take ratios of these to cancel out the dependence upon the unwanted $p_A(1-p_A)$. Their approach follows the layout of an **analysis of variance** table, although it is not a conventional analysis of variance model. The idea of partitioning variances in this layout is sometimes called ***Analysis of Molecular Variance***, or ***AMOVA***.

To estimate F and θ, we need data from at least two replicate populations: for example, contemporary populations on two different islands. The data that we can observe are sample allele frequencies and correlations within each of the populations. Following Weir (1996), we use the symbol tilde ($\widetilde{}$) to denote a sample frequency. For example, if we sample $2n_i$ alleles from population i, then

$$\widetilde{p}_{Ai} = \frac{\text{number of the } 2n_i \text{ alleles that are of type } A}{2n_i}\,.$$

More precisely:

- there are r replicate populations: populations $i = 1, 2, \ldots, r$;
- there are n_i individuals sampled from population i: individuals $j = 1, 2, \ldots, n_i$;
- each individual has 2 alleles: alleles $k = 1, 2$.

For allele k of individual j in population i, define the indicator

$$Y_{ijk} = \begin{cases} 1 & \text{if allele } ijk \text{ is of type } A, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\widetilde{p}_{Ai} = \frac{1}{2n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{2} Y_{ijk} = \overline{Y_i}$$

Similarly, the sample frequency of homozygotes for allele type A is

$$\widetilde{\mathcal{P}}_{AAi} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij1}Y_{ij2}$$

Clearly, $\mathbb{E}(\widetilde{p}_{Ai}) = p_A$, and $\mathbb{E}(\widetilde{\mathcal{P}}_{AAi}) = \mathcal{P}_{AA}$, where the expectation is taken over all replicate populations. We will also need $\mathbb{E}(\widetilde{p}_{Ai}^2)$, which we find by computing $\mathrm{var}(\widetilde{p}_{Ai})$:

Result: Consider the process of drawing a replicate population (genetic sampling), then taking a sample of size n_i (statistical sampling) and finding $\widetilde{p}_{Ai}$. The variance of $\widetilde{p}_{Ai}$ is

$$\begin{aligned}\mathrm{var}(\widetilde{p}_{Ai}) = \mathbb{E}(\widetilde{p}_{Ai}^2) - p_A^2 &= \mathcal{P}_{A/A} - p_A^2 + \frac{1}{2n_i}\left(p_A + \mathcal{P}_{AA} - 2\mathcal{P}_{A/A}\right) \\ &= p_A(1-p_A)\left\{\theta + \frac{1}{2n_i}\left(1 + F - 2\theta\right)\right\},\end{aligned}$$

where expectations are taken over the entire process of genetic sampling and statistical sampling. This is called the *total variance of allele frequency* by Weir (1996), p. 48, eqn 2.14.

Proof: Consider

$$\widetilde{p}_{Ai} = \frac{1}{2n_i}\sum_{j=1}^{n_i}(Y_{ij1} + Y_{ij2}) = \frac{1}{2n_i}\sum_{j=1}^{n_i} X_j\,,$$

where $X_j = Y_{ij1} + Y_{ij2}$. Using the standard expression for the variance of a sum:

$$\mathrm{var}(\widetilde{p}_{Ai}) = \frac{1}{4n_i^2}\left\{\, n_i\mathrm{var}(X_j) + n_i(n_i-1)\mathrm{cov}(X_j, X_\ell)\right\} \qquad (\ell \neq j)\,. \tag{10}$$

Now

$$\begin{aligned}\mathrm{var}(X_j) = \mathrm{var}\left(Y_{ij1} + Y_{ij2}\right) &= 2\mathrm{var}(Y_{ijk}) + 2\mathrm{cov}\left(Y_{ij1}, Y_{ij2}\right) \\ &= 2\left(p_A - p_A^2\right) + 2\left(\mathcal{P}_{AA} - p_A^2\right) \\ &= 2(p_A + \mathcal{P}_{AA} - 2p_A^2)\,,\end{aligned}$$

noting that $\mathbb{E}(Y_{ijk}) = p_A = \mathbb{E}(Y_{ijk}^2)$, and $\mathbb{E}(Y_{ij1}Y_{ij2}) = \mathcal{P}_{AA}$, so $\mathrm{var}(Y_{ijk}) = p_A - p_A^2$ and $\mathrm{cov}(Y_{ij1}, Y_{ij2}) = \mathcal{P}_{AA} - p_A^2$.

Also,

$$\begin{aligned}\mathrm{cov}(X_j, X_\ell) &= \mathbb{E}\left\{\left(Y_{ij1} + Y_{ij2}\right)\left(Y_{i\ell1} + Y_{i\ell2}\right)\right\} - \mathbb{E}\left(Y_{ij1} + Y_{ij2}\right)\mathbb{E}\left(Y_{i\ell1} + Y_{i\ell2}\right) \\ &= 4\,\mathrm{cov}\left(Y_{ijk}, Y_{i\ell m}\right) \qquad (\ell \neq j) \\ &= 4\left(\mathcal{P}_{A/A} - p_A^2\right)\,,\end{aligned}$$

because $\mathbb{E}(Y_{ijk}Y_{i\ell m}) = \mathcal{P}_{A/A}$ for $\ell \neq j$. Inserting these expressions into (10) gives the result. □

Weir and Cockerham's estimators

The data observed from real populations can be written as y_{ijk}, the observed indicator for whether allele k in individual j in population i is of type A. Weir and Cockerham's idea is

to partition the overall sum of squares of the $\{y_{ijk}\}$ into a hierarchy of intermediate sums of squares: within individuals; between individuals within populations; and between populations:

$$\begin{aligned}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(y_{ijk}-\overline{y})^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}\Big\{(y_{ijk}-\overline{y}_{ij})+(\overline{y}_{ij}-\overline{y}_i)+(\overline{y}_i-\overline{y})\Big\}^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(y_{ijk}-\overline{y}_{ij})^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(\overline{y}_{ij}-\overline{y}_i)^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(\overline{y}_i-\overline{y})^2\,,\end{aligned} \tag{11}$$

where

$$\overline{y}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}y_{ijk}}{2n.}\,,\qquad \overline{y}_i=\frac{\sum_{j=1}^{n_i}\sum_{k=1}^{2}y_{ijk}}{2n_i}\,,\qquad \overline{y}_{ij}=\frac{\sum_{k=1}^{2}y_{ijk}}{2}\,,\qquad n.=\sum_{i=1}^{r}n_i$$

Equation (11) results because the cross-terms in each sum of squares vanish. Following a lot of tedious algebra, the following expressions emerge, equivalent to Weir (1996), p. 177, Table 5.4:

- Level 1 (alleles within individuals). This level is given code **G** (for Gametes).

Simplification: $$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(y_{ijk}-\overline{y}_{ij})^2=\sum_{i=1}^{r}n_i\left(\widetilde{p}_{Ai}-\widetilde{\mathcal{P}}_{AAi}\right)$$

Expectation: $$\mathbb{E}\left\{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(y_{ijk}-\overline{y}_{ij})^2\right\}=\mathbb{E}\left\{\sum_{i=1}^{r}n_i\left(\widetilde{p}_{Ai}-\widetilde{\mathcal{P}}_{AAi}\right)\right\}=n.p_A(1-p_A)(1-F)$$

- Level 2 (**I:** individuals within populations).

$$\begin{aligned}\text{Simplification:}\quad \textstyle\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(\overline{y}_{ij}-\overline{y}_i)^2 &= \sum_{i=1}^{r}n_i\left(\widetilde{p}_{Ai}+\widetilde{\mathcal{P}}_{AAi}-2\,\widetilde{p}_{Ai}^{\,2}\right)\\ \text{Expectation:}\quad \mathbb{E}\left\{\textstyle\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(\overline{y}_{ij}-\overline{y}_i)^2\right\} &= \mathbb{E}\left\{\sum_{i=1}^{r}n_i\left(\widetilde{p}_{Ai}+\widetilde{\mathcal{P}}_{AAi}-2\,\widetilde{p}_{Ai}^{\,2}\right)\right\}\\ &= p_A(1-p_A)(n.-r)\Big\{1-F+2(F-\theta)\Big\}\end{aligned}$$

- Level 3 (**P:** between Populations):

$$\begin{aligned}\text{Simplification:}\quad \textstyle\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(\overline{y}_i-\overline{y})^2 &= \sum_{i=1}^{r}2n_i\left(\widetilde{p}_{Ai}-\widetilde{p}_A\right)^2 \qquad (\widetilde{p}_A=\overline{y})\\ \text{Expectation:}\quad \mathbb{E}\left\{\textstyle\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{2}(\overline{y}_i-\overline{y})^2\right\} &= \mathbb{E}\left\{\sum_{i=1}^{r}2n_i\left(\widetilde{p}_{Ai}-\widetilde{p}_A\right)^2\right\}\\ &= p_A(1-p_A)(r-1)\Big\{1-F+2(F-\theta)+2n_c\,\theta\Big\}\end{aligned}$$

where $n_c = \frac{1}{r-1}\left(n_{\cdot} - \frac{\sum_{i=1}^{r} n_i^2}{n_{\cdot}}\right)$ acts as an average of the sample sizes, and emerges from the algebra when finding the expectations. The term n_c indicates that it acts as a 'combined' sample size per population.

Overall, the working produces Table 1, which is written in an analysis-of-variance type format, and corresponds to Table 5.4 on page 177 of Weir (1996).

To finish off, we use the method-of-moments to find $\widehat{\theta}$ and $\widehat{F}$ as follows (see Table 1 for notation):

$$\begin{aligned}
\mathbb{E}(\text{MSP}) &= p_A(1-p_A)\Big\{1 - F + 2(F-\theta) + 2n_c\,\theta\Big\} \\
\mathbb{E}(\text{MSI}) &= p_A(1-p_A)\Big\{1 - F + 2(F-\theta)\Big\} \\
\mathbb{E}(\text{MSG}) &= p_A(1-p_A)\Big\{1 - F\Big\}
\end{aligned}$$

Manipulating gives:

$$\theta = \frac{\mathbb{E}(\text{MSP}) - \mathbb{E}(\text{MSI})}{\mathbb{E}(\text{MSP}) + (n_c-1)\mathbb{E}(\text{MSI}) + n_c\mathbb{E}(\text{MSG})} \quad \Rightarrow \quad \widehat{\theta} = \frac{\text{MSP} - \text{MSI}}{\text{MSP} + (n_c-1)\text{MSI} + n_c\,\text{MSG}}\,; \tag{12}$$

$$F = 1 - \frac{2n_c\mathbb{E}(\text{MSG})}{\mathbb{E}(\text{MSP}) + (n_c-1)\mathbb{E}(\text{MSI}) + n_c\mathbb{E}(\text{MSG})} \quad \Rightarrow \quad \widehat{F} = 1 - \frac{2n_c\text{MSG}}{\text{MSP} + (n_c-1)\text{MSI} + n_c\,\text{MSG}}\,; \tag{13}$$

$$f = \frac{F-\theta}{1-\theta} \quad \Rightarrow \quad \widehat{f} = \frac{\widehat{F}-\widehat{\theta}}{1-\widehat{\theta}}\,. \tag{14}$$

Note that these expressions are not exactly unbiased for θ and F, because the expected ratio is not equal to the ratio of expectations. However, they have the big advantage of avoiding altogether the need to estimate p_A, the proportion of allele type A in the infinite ancestral population that probably never existed. They are entirely compiled from observable data from populations $i = 1, 2, \ldots, r$.

Combining estimates for different alleles and loci

The estimators (12), (13), and (14) are gained from a single allele type, A, at a single locus. In practice, there will be many loci, each with many different allele types. Each allele type will give a different *estimate* of θ, F, and f; but all alleles are estimating the same *parameter value*. We therefore need a way of combining the estimates from different allele types and loci into one single estimate from a data set.

There is also a subtlety here regarding the sample size, n_i, from population i. For one particular locus, n_i is the number of individuals successfully genotyped at that locus, from population i. All alleles at the locus therefore have the same sample size n_i. However, for different loci, the sample sizes for population i may differ, because some individuals will genotype successfully

at some loci but not at others. For example, at locus 1 we might have $n_i = 31$ successful genotypes, but at locus 2 we might have only $n_i = 28$.

There is no 'right' way to combine the allele-specific estimates into an overall estimate, so various authors have suggested various schemes which are tested by simulation. The ideal scheme will lead to good properties with respect to both accuracy and precision.

Consider the estimates from allele type m at locus ℓ. In equations (12) and (13), we could write

$$\widehat{\theta}_{\ell m} = \frac{\texttt{theta.top}_{\ell m}}{\texttt{denom}_{\ell m}}\,, \qquad \widehat{F}_{\ell m} = 1 - \frac{\texttt{F.top}_{\ell m}}{\texttt{denom}_{\ell m}}\,.$$

Our need is to combine the multiple estimates for all ℓ and m into a single estimate. Weir (1996) suggests summing the top terms across alleles, summing the denominator across alleles, and dividing so that the overall estimate is a ratio of means rather than a mean of ratios. However, Weir suggests a by-locus weighting that would give equal weight to all loci, regardless of how many individuals were successfully genotyped at each locus. Genepop on the Web broadly uses Weir's approach, but they use a different weighting such that loci with fewer individuals genotyped will contribute less to the overall estimates than loci with more individuals genotyped: see `http://genepop.curtin.edu.au/Appendix2.htm`. In practice, the difference tends to be very small because the sample sizes do not differ greatly between loci. We use the Genepop method here.

The overall estimates of θ and F from multiple alleles and loci are therefore as follows, where sums are taken over loci ℓ and over allele types within loci, m:

$$\widehat{\theta} = \frac{\sum_\ell \sum_m \texttt{theta.top}_{\ell m}}{\sum_\ell \sum_m \texttt{denom}_{\ell m}} = \frac{\sum_\ell \sum_m \{\mathrm{MSP}_{\ell m} - \mathrm{MSI}_{\ell m}\}}{\sum_\ell \sum_m \{\mathrm{MSP}_{\ell m} + (n_{c,\ell m} - 1)\mathrm{MSI}_{\ell m} + n_{c,\ell m}\,\mathrm{MSG}_{\ell m}\}}\,. \tag{15}$$

$$\widehat{F} = 1 - \frac{\sum_\ell \sum_m \texttt{F.top}_{\ell m}}{\sum_\ell \sum_m \texttt{denom}_{\ell m}} = 1 - \frac{\sum_\ell \sum_m \{2\,n_{c,\ell m}\,\mathrm{MSG}_{\ell m}\}}{\sum_\ell \sum_m \{\mathrm{MSP}_{\ell m} + (n_{c,\ell m} - 1)\mathrm{MSI}_{\ell m} + n_{c,\ell m}\,\mathrm{MSG}_{\ell m}\}}\,. \tag{16}$$

These expressions are equivalent to the calculations done on Genepop on the Web (Raymond and Rousset, 1995; Rousset, 2008), and in the function `Fstatistics.func` in the R bundle. The within-population inbreeding coefficient f is estimated from $\widehat{f} = \frac{\widehat{F} - \widehat{\theta}}{1 - \widehat{\theta}}$ using the overall $\widehat{F}$ and $\widehat{\theta}$ estimates.

Haploid data

In the case of haploid data (e.g. mtDNA), the inbreeding coefficients F and f are no longer relevant, but $\theta = F_{ST}$ is still relevant and commonly used to measure population structure and gene flow from haplotypes. Instead of indicators y_{ijk}, we do not have the level of alleles within

	Source	d.f.	Sum of Squares	Mean Square	Expected Mean Square
P:	Between populations	$r-1$	$\sum_{i=1}^{r} 2n_i \left(\widetilde{p}_{Ai} - \widetilde{p}_A\right)^2$	$\text{MSP} = \dfrac{\sum_{i=1}^{r} 2n_i \left(\widetilde{p}_{Ai} - \widetilde{p}_A\right)^2}{r-1}$	$\mathbb{E}(\text{MSP}) = p_A(1-p_A)\left\{1 - F + 2(F-\theta) + 2n_c\,\theta\right\}$
I:	Individuals within populations	$n_\cdot - r$	$\sum_{i=1}^{r} n_i \left(\widetilde{p}_{Ai} + \widetilde{\mathcal{P}}_{AAi} - 2\,\widetilde{p}_{Ai}^{\,2}\right)$	$\text{MSI} = \dfrac{\sum_{i=1}^{r} n_i \left(\widetilde{p}_{Ai} + \widetilde{\mathcal{P}}_{AAi} - 2\,\widetilde{p}_{Ai}^{\,2}\right)}{n_\cdot - r}$	$\mathbb{E}(\text{MSI}) = p_A(1-p_A)\left\{1 - F + 2(F-\theta)\right\}$
G:	Alleles within individuals	$n_\cdot$	$\sum_{i=1}^{r} n_i \left(\widetilde{p}_{Ai} - \widetilde{\mathcal{P}}_{AAi}\right)$	$\text{MSG} = \dfrac{\sum_{i=1}^{r} n_i \left(\widetilde{p}_{Ai} - \widetilde{\mathcal{P}}_{AAi}\right)}{n_\cdot}$	$\mathbb{E}(\text{MSG}) = p_A(1-p_A)\left\{1 - F\right\}$

Table 1: Hierarchical levels of variance written in an ANOVA format, as in Table 5.4 on page 177 of Weir (1996). The 'combined sample size' per population is defined as $n_c = \frac{1}{r-1}\left(n_\cdot - \frac{\sum_{i=1}^{r} n_i^2}{n_\cdot}\right)$. The quantities shown in the table are suitable for estimating F and θ for diploid data.

individuals. The observations are indicators y_{ij} for whether individual j in population i has haplotype A. The partitioning of sums of squares corresponding to (11) becomes

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\overline{y}_i-\overline{y})^2\,.$$

Following the same scheme as for diploid data, the lengthy algebra finally results in

$$\widehat{\theta}_{\text{haploid}} = \frac{\text{MSP}_{\text{hap}} - \text{MSG}_{\text{hap}}}{\text{MSP}_{\text{hap}} + (n_c-1)\text{MSG}_{\text{hap}}}\,,$$

where $n_c = \frac{1}{r-1}\left(n_{\cdot} - \frac{\sum_{i=1}^{r} n_i^2}{n_{\cdot}}\right)$ as before, $\text{MSP}_{\text{hap}} = \frac{\sum_{i=1}^{r} n_i(\widetilde{p}_{Ai}-\widetilde{p}_A)^2}{r-1}$ and $\text{MSG}_{\text{hap}} = \frac{\sum_{i=1}^{r} n_i\widetilde{p}_{Ai}(1-\widetilde{p}_{Ai})}{n_{\cdot}-r}$. See Weir (1996), Table 5.3, page 172.

AMOVA: Inference about population structure from data

The idea of AMOVA (Analysis of Molecular Variance) is to compare the variability drawn from different sources in the ANOVA-like layout in Table 1, and to draw inference about $\theta = F_{ST}$ through numerical resampling procedures such as permutation tests and bootstrapping.

Although some of the ideas apply to F and f as well, if inbreeding is of interest it is more likely to be tested using within-population tests for Hardy-Weinberg proportions: noting that f describes the deviation of genotypes from HW proportions (page 21). We will focus on $F_{ST} = \theta$ here, as it is used to describe population differentiation or structure.

Bootstrapped confidence intervals

The alleles within a locus are not statistically independent, because having more of one allele type necessitates having less of another. Therefore, different alleles within a locus do not provide statistically independent estimates of θ. For an extreme example, if there are only two allele types at a locus, it is easy to show that the estimate in equation (12) will be the same for both alleles.

However, most studies in population genetics choose loci on different chromosomes (*unlinked loci*), for example a suite of microsatellite loci is typically chosen to ensure this. If so, then loci are statistically independent and estimates of θ from different loci are statistically independent. In this case, confidence intervals can be gained for θ by bootstrapping the data across loci. However, many loci are needed for this to be effective. In practice, we may be more interested in specific hypotheses such as $H_0 : \theta = 0$, and these are more effectively tested by permutation tests.

AMOVA permutation test for $H_0 : \theta = 0$.

Permutation tests are described briefly in Weir (1996), and in the manual to the population genetics software Arlequin (Excoffier, Laval, and Schneider, 2006): Section 7.2, p.119 - 127.

Under the null hypothesis $H_0 : \theta = 0$, there is no population differentiation, so any individual could be drawn from any population. The permutation test works by permuting the population

labels such that the data for each individual is assigned to a randomly generated population, keeping the sample sizes constant for each population.

In R, if the data frame `dat` contains `ndot` individuals, and their population labels are in `dat$pop`, we can achieve the necessary permutation as follows:

```
> dat.perm = dat
> dat.perm$pop = sample(dat$pop, ndot, replace=F)
```

Then θ is estimated for the permuted data in `dat.perm`. This process is repeated many times (say, 500 times) to generate 500 estimates of θ under the null hypothesis. If the real-data value of θ is greater than 95% of the simulated values, it is considered unusually large for the null distribution and there is evidence that the true value of θ is not 0. The p-value for the one-sided test would be:

$$p = \frac{\text{number of the 500 sims giving estimated } \theta \text{ greater than the real-data estimate of } \theta}{500}\,.$$

A one-sided test is used because (as a general rule) the true value of θ is ≥ 0. For example, θ can be defined as a *probability* of identity-by-descent; although we note that under the more general definition as a correlation θ could in principle be < 0.

Fixed populations versus random populations

Throughout our development, we have been treating populations as ***random replicate populations*** and using observed data to deduce properties of the mechanism for generating these random populations. Our inference and interest is therefore about the ***population-generating mechanism***, not about the specific populations we happen to observe.

We could alternatively imagine that our populations are fixed and that these fixed populations are our object of interest and inference. The statistic $F_{ST} = \widehat{\theta}$ could be calculated using equation (15) just as before. However, as Weir (1996, page 166-167) points out, there are more direct ways of determining differentiation between fixed populations, for example using a chi-squared test for equality of their allele frequencies. If they have different allele frequencies, they may be considered as 'differentiated' populations.

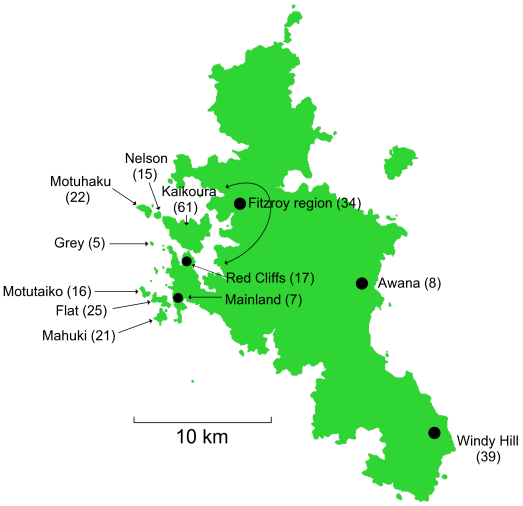
Example: ship rats (*Rattus rattus*) on Great Barrier Island, New Zealand

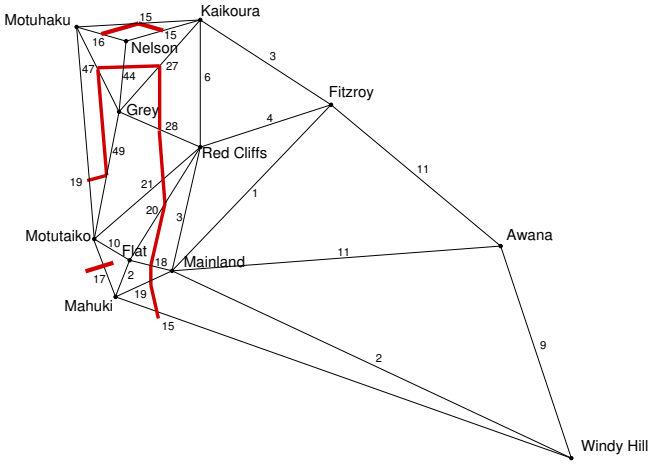
We illustrate here how $\theta = F_{ST}$ can be used to get a rough picture of landscape barriers to gene flow, using data from ship rats (*Rattus rattus*) on Great Barrier Island, New Zealand. The analysis below is crude and simple, but does appear to provide some insight into barriers to ship rat dispersal.

Code	Awa	BM	Fit	Flat	Grey	Haku	Kai	Mahu
Population	Awana	Mainland	Fitzroy	Flat	Grey	Motuhaku	Kaikoura	Mahuki
n_i	8	7	34	25	5	22	61	21

Code	Nel	Red	Taik	WH
Population	Nelson	Red Cliffs	Motutaiko	Windy Hill
n_i	15	17	16	39

The F_{ST} network analysis is described in Fewster, Miller, and Ritchie (2011). The network is created using a Delaunay triangulation given the coordinates of each location. $F_{ST} = \theta$ is estimated separately for every pair of populations in the network. Boundaries are grown using the Monmonier algorithm (Monmonier, 1973) by picking the edge with the largest F_{ST}, and growing the boundary until it can no longer find an edge with F_{ST} above a pre-set threshold. Here, the threshold used is $F_{ST} = 0.15$.





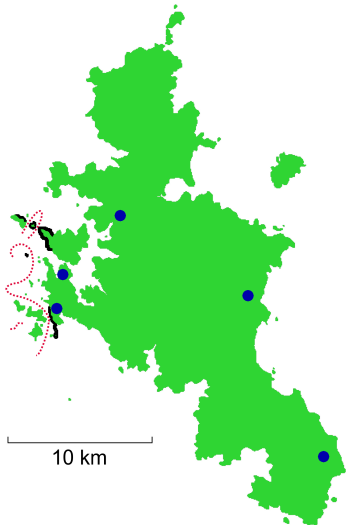


Figure 8: F_{ST} network for ship rats on Great Barrier Island, New Zealand. F_{ST} estimates are multiplied by 100 on the network. The geographical locations of boundaries are marked on the map on the right. Locations of cliffs are shown in bold lines on the map.

A rule of thumb for interpretation of $\theta = F_{ST}$ values is given by Wright (1978), although it should be stressed that it is only a very crude guideline:

- $F_{ST} = 0$ to 0.05 : ***little*** genetic differentiation.
- $F_{ST} = 0.05$ to 0.15 : ***moderate*** genetic differentiation.
- $F_{ST} = 0.15$ to 0.25 : ***great*** genetic differentiation.
- $F_{ST} = 0.25$ to 1 : ***very great*** genetic differentiation.

For the ship rat example, there is an association between high values of F_{ST} and either cliffs or long water crossings (> 1 km). There is relatively little genetic distinction between locations separated by much longer distances over land (20-30 km), because there are no significant habitat barriers to rat dispersal. Interestingly, the link between Kaikoura Island and Fitzroy is stronger than that between Kaikoura and Red Cliffs, even though the separation distance is much smaller between the latter pair. This is probably due to a combination of fast currents through the dividing channel between Kaikoura and Red Cliffs, and the fact that Fitzroy is the hub for boat traffic to and from Kaikoura.

The linking of genetic boundaries to landscape features is described as *landscape genetics*.

Mantel test for isolation-by-distance

A popular way of using F_{ST} estimates is to conduct a test of isolation-by-distance, in which pairwise F_{ST} estimates are related to pairwise geographical distances. A Mantel test is a way of testing for correlations between two matrices: for example, whether genetic distance is correlated with geographical distance. Like the other tests, it is based on permutations. Rows and columns of the first matrix are permuted, and after each permutation, the matrix correlation between the permuted first matrix and the second matrix is computed. The correlation between the real-data matrices is then compared against this null distribution to test for significance of the correlation.

In our context, we have the matrix D such that D_{ij} is the geographical distance between sampled locations i and j; and the matrix G of 'genetic distances', where G_{ij} might be the pairwise F_{ST} between locations i and j, or some transformation of F_{ST} such as $F_{ST}/(1 - F_{ST})$ (suggested in Genepop on the Web); or $(1/F_{ST} - 1)/4$ corresponding to Nm in the infinite island model (see Bohonak, 2002); or some other genetic distance, such as distances based on the number of shared alleles among individuals. If the test is significant, there is evidence that the genetic separation of populations is correlated with their geographical separation. Other geographical 'distances' that incorporate habitat type can also be used.

Practical Session 3

Open the file `PopGen-Practical-3.txt` and follow the instructions within to do the following examples.

***Exercise* 7:** *Answers on page 43.*

(a) Estimate $\theta = F_{ST}$ for the Great Barrier Island data using R. Only use individuals with at least 6 loci successfully genotyped:
`res = Fstatistics.func(gbi.dat, min.loci.per.indiv=6)`
Does the overall θ estimate indicate much population differentiation?

(b) Check that the R code gives the same answers for $\theta = F_{ST}$ as Genepop on the Web.

(c) Look at the pairwise F_{ST} estimates in `res$pairs`. Check a few of them against the network in Figure 8 to be sure they are the same.

(d) Write your own code in R to do an AMOVA permutation test of the hypothesis $H_0 : \theta = 0$ for the overall θ on all populations. Is there significant evidence that $\theta > 0$?
Note: use `Fstatistics.func` with the option `allpairs=F` to save time in the permutation test. The option `allpairs=F` means that only the overall θ will be calculated for each permutation; not all pairwise θ values.

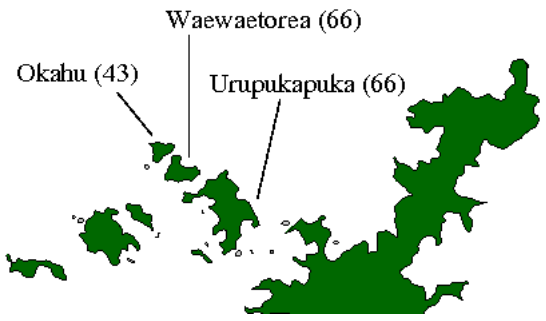
(e) Do a pairwise AMOVA permutation test for θ between the two populations 'Mainland' (BM) and Fitzroy (Fit). The easiest way is to create a new data set containing just these two populations, and use the code you wrote for (d):

```
new.dat = gbi.dat[gbi.dat$pop=="BM" | gbi.dat$pop=="Fit", ]
```

Is the differentiation between these populations significant? Comment on this in the light of the sample sizes from the two populations: 7 from BM, 34 from Fit.

Note: You can check your answer for (e) using the function
`perm.theta.func(new.dat, Nsim=500, allpairs=F)`
Increase `Nsim` for a more accurate p-value.

(f) Another data set of Norway rats (*Rattus norvegicus*) from three islands in the Bay of Islands, New Zealand, is also provided in `boi.dat`. Find $F_{ST} = \theta$, and use `perm.theta.func(boi.dat)` to test $H_0 : \theta = 0$.



Under the assumptions of the Infinite Island Model, what is the effective number of migrants per generation in each of the three populations?
Note: this number is not usually interpreted literally, but rather as a comparative measure.

Effective population size, N_e

In the models we have examined, the rate of change of genetic quantities from one generation to the next depends fundamentally on the population size N. Conversely, information about the rate of change of genetic quantities, such as allele frequencies or inbreeding, should provide information about population size N. However, all the models are highly idealised and do not necessarily describe the reproduction of real animals. In practice, individuals are not equally successful as breeders, and some contribute many more gametes to the next generation than others. Although our models do allow for some variance in the number of gametes contributed across individuals, the variance in our idealised models is much lower than it is likely to be in reality. For this reason, the notion of *genetic effective population size, N_e,* has been developed. The effective population size is the size of an idealised population whose genetic parameters are changing at the same rate as those in the population we have observed. The *ideal* population meets the three conditions of equal sex ratio, random mating, and constant census population size over generations. (Note that we have not looked at models that include sexes: the ideal population is a dioecious version of the Wright-Fisher model.) The idea is that the real population, with a census size of N_c individuals, can then be studied *as if it were an ideal population with size N_e individuals.*

Generally, because of the uneven contribution of individuals to gametes, the effective population size N_e is smaller than the census population size N_c, and often it is much smaller (perhaps 10 or even 100 times smaller, depending on the species and mating system). Additionally, the ratio of N_e to N_c is not constant or predictable, and there are several possible definitions of N_e depending upon which genetic parameters are inspected. Although N_e is a parameter of fundamental importance in evolutionary genetics, it is less clear how useful or relevant it is for describing population size of a contemporary population, or for use in conservation or management. Some discussion and references are provided in Russell and Fewster (2009).

Solutions to Exercises

Exercise 1. $1-\theta_t$ is the probability that two different alleles a_1 and a_2 in generation t are *not* IBD. Trace back the ancestors of alleles a_1 and a_2 for generations $t-1, t-2, \ldots, 1, 0$. If a_1 and a_2 share the same ancestor in *any* of these generations, they are IBD. In each generation, the probability that they share the same ancestor is $1/(2N)$. There are t generations between 0 and $t-1$, each giving an opportunity for sharing the same parent.

Thus, to *not* be IBD, all t of the opportunities must fail. So $(1-\theta_t) = \left(1 - \dfrac{1}{2N}\right)^t$.

Exercise 2.

(a) The only way to reduce $\theta = F_{ST}$ for a fixed t is to increase N. Neither `p.ancestral` nor `nrep` will have any effect. This can also be deduced by $\theta = 1 - \left(1 - \frac{1}{2N}\right)^t$.

(b) The only way to increase $\theta = F_{ST}$ when N is fixed is to increase t. The differentiation between populations increases with the number of generations for which they have been separated.

(c) The scatter of the grey points is increased by decreasing `nrep`. This is a consequence of increasing simulation variability: it is nothing to do with the genetic process.

(d) The deviation of the coloured lines from $1 - \left(1 - \frac{1}{2N}\right)^t$ is governed by `p.ancestral`. The coloured lines $\frac{\mathcal{P}_{A/A} - p_A^2}{p_A(1-p_A)}$ will have high variance if the denominator $p_A(1-p_A)$ is small. The most precise results use `p.ancestral=c(0.5, 0.5)`, whereas `p.ancestral=c(0.01, 0.99)` gives much less precise estimates. Like (c), this is not connected with the genetic process itself, but it does suggest that the ancestral allele frequencies may affect the precision of *estimates* of θ.

Exercise 3.

(a) The conditional distribution is $Y_{t+1} \,|\, Y_t \sim \text{Binomial}\left(2N, \frac{Y_t}{2N}\right)$. Thus

$$p_{ij} = \mathbb{P}(Y_{t+1} = j \,|\, Y_t = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}.$$

(b) From the transition probabilities in (a), we can see that the Markov chain is irreducible and aperiodic, and it has a finite state space. Therefore, an equilibrium distribution exists, and the chain will converge to the distribution. The possible equilibrium states are the two absorbing states, $Y_t = 0$ and $Y_t = 2N$.

(c) Using the Binomial distribution in (a),

$$\mathbb{E}(Y_{t+1} \,|\, Y_t) = 2N \times \frac{Y_t}{2N} = Y_t\,, \text{ as required.}$$

(d) Generation 0 is drawn from the infinite reference population, so $Y_0 \sim \text{Binomial}(2N, p_A)$, and $\mathbb{E}(Y_0) = 2Np_A$. At the stopping time T, the only states that Y_T can take are 0 or 2N:

$$Y_T = \begin{cases} 2N & \text{with probability } \alpha, \\ 0 & \text{with probability } 1-\alpha. \end{cases}$$

Thus $\mathbb{E}(Y_T) = 2N\alpha + 0(1-\alpha) = 2N\alpha$.

Using $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$, we have

$$2N\alpha = 2Np_A \quad \Rightarrow \quad \alpha = p_A.$$

Thus the probability that the population is eventually fixed for allele A is the same as the frequency of A in the infinite reference population.

Exercise 4. Using the expressions $H_T = 2p_A(1-p_A)$, $H_S = 2(p_A - \mathcal{P}_{A/A})$, and $H_I = 2(p_A - \mathcal{P}_{AA})$ from page 23, the results follow immediately.

Exercise 5.

(a) In the Wright-Fisher model, θ depends only upon t and N. Migration in the infinite island model lowers θ relative to the Wright-Fisher model. Therefore, for a fixed t, increasing N in the W-F model is the only way the W-F θ can be changed to match θ in the infinite island model. For any level of migration m, we can always find a large enough N to compensate for one particular t, but not for all t because the W-F θ will always converge to 1 eventually.

(b) For a given N and t, the Wright-Fisher θ will always be higher than the infinite island θ, so the infinite island $\theta_t < 1 - \left(1 - \frac{1}{2N}\right)^t$ for all t. Thus it cannot converge to equilibrium $\frac{1}{4Nm+1}$ until $1 - \left(1 - \frac{1}{2N}\right)^t \geq \frac{1}{4Nm+1}$. As m decreases, the right-hand side increases, so a larger t is needed before convergence can occur.

(c) With $m = 1$, all alleles are drawn directly from the infinite reference population, so $\mathcal{P}_{A/A} = p_A^2$ and $\theta = 0$.
The $\frac{1}{4Nm+1}$ formula is not correct because it is an approximation suitable for small m, not for $m = 1$. The exact expression is $\theta_\infty = (1-m)^2 / \{2N - (1-m)^2(2N-1)\}$ which is 0 when $m = 1$.

Exercise 6.

(a) $f = \frac{F-\theta}{1-\theta} = 0$, so $F = \theta$. When $f = 0$, the inbreeding model is identical to the Wright-Fisher model.

(b) Only f can be estimated from data on a single population, as it is a within-population parameter. Both θ and F pertain to the relationship between replicate populations, so cannot (or should not) be estimated from a single population.

(c) Yes, the individual is more likely to be a homozygote than one sampled directly from the infinite reference population, despite having $f = 0$. This is because of the effect of population division, which causes an increase in overall 'inbreeding' even though there is no inbreeding effect within any of the individual populations. This is the meaning of the parameter F, the overall inbreeding parameter. If $f = 0$, then $F = \theta$, so $F = \frac{\mathcal{P}_{AA} - p_A^2}{p_A(1-p_A)}$ is greater than 0, ensuring that homozygotes (corresponding to $\mathcal{P}_{AA}$) are more common in the overall population than they would be under Hardy-Weinberg proportions (corresponding to p_A^2).

(d) First type `attach(res)` so the components of res can be accessed directly.

(i) Verify that $\mathbb{E}_\star(p_{Ai\star}^2) = \mathcal{P}_{A/A}$:
For each allele type (1 or 2), we need to compare `pAsq` at time $t-1$ with `PA.A` at time t. Thus we link $t-1$ (1:9) for `E(pA^2)` with t (2:10) for `PA.A`, and observe that the points lie on the $y = x$ red line: `plot(pAsq[1, 1:9], PA.A[1, 2:10]);abline(0, 1, col=2)`
Same again for allele type 2: `plot(pAsq[2, 1:9], PA.A[2, 2:10]);abline(0, 1, col=2)`
It doesn't work to link `E(pA^2)` with `PA.A` at the same generation:
`plot(pAsq[1, ], PA.A[1, ]);abline(0, 1, col=2)`

(ii) Verify that θ is the percentage of variance between populations: again, we need the parental generation when computing the variance $\mathbb{E}_\star(p_{Ai\star}^2) - \mathbb{E}_\star(p_{Ai\star})^2$. The overall variance is 0.4×0.6 for these alleles. For θ, the best object to use is `two.indiv.ibd`. Alternatively, we can use `theta.by.allele.type[1,]` or `theta.by.allele.type[2,]`.

```
plot((pAsq[1, 1:9] - pA[1, 1:9]^2)/(0.4*0.6), two.indiv.ibd[2:10])
abline(0, 1, col=2)
## or:
plot((pAsq[1, 1:9] - pA[1, 1:9]^2)/(0.4*0.6), theta.by.allele.type[1, 2:10])
abline(0, 1, col=2)

## For allele type 2:
plot((pAsq[2, 1:9] - pA[2, 1:9]^2)/(0.4*0.6), theta.by.allele.type[2, 2:10])
abline(0, 1, col=2)
```

(iii) We do not expect that $\mathbb{E}_i(p_{Ai}^2) = \mathcal{P}_{AA}$, either at same generation or different generations, because $\mathcal{P}_{AA}$ incorporates the extra effect of within-population inbreeding. The following plots do not lie on the red lines.

```
plot(pAsq[1, ], PAA[1,]); abline(0, 1, col=2)
points(pAsq[1, 1:9], PAA[1, 2:10], col=3)
```

Exercise 7.

(a) The overall θ estimate is 0.130. Using the rule of thumb on page 37, this is moderate to high population differentiation.

(d) Possible code for the permutation test is below. The code can be pasted from the PDF into *R*.

```
## Find the real-data value of theta:
real.theta = Fstatistics.func(gbi.dat, min.loci.per.indiv=6, allpairs=F)["theta"]
## Set up vector to store permutation values of theta for 500 permutations:
perm.res = rep(NA, 500)
n.indiv = nrow(gbi.dat)
## Run the permutations:
for(sim in 1:500){
  perm.dat = gbi.dat
  ## Permute the population labels:
  perm.dat$pop = sample(perm.dat$pop, size=n.indiv, replace=F)
  ## Find theta for the permuted data, and insert into perm.res:
  perm.res[sim]=Fstatistics.func(perm.dat,min.loci.per.indiv=6,allpairs=F)["theta"]
}
## View histogram of the results: ensure the x-axis can display the real value too:
hist(perm.res, col="blue", xlim=range(c(perm.res, real.theta)))
## Draw a vertical line where the real data result lies:
abline(v=real.theta, col=2, lwd=2)
## Find the p-value:
length(perm.res[perm.res > real.theta]) / 500
```

(e) The p-value for the single pair BM and Fitzroy should be close to 0.15. Thus there is no evidence against $H_0 : \theta = 0$ for this pair of populations. The estimated θ is very small (0.012). However, the sample size from BM is very small (7 individuals), and this will contribute to the non-significant result.

(f) Using `perm.theta.func(boi.dat)` shows very small, but statistically significant, values of θ overall and between each pair of populations. The overall θ is 0.024. Under the Infinite Island Model assumptions, this gives

$$0.024 = \frac{1}{4Nm + 1} \quad \Rightarrow \quad Nm = 10.1.$$

The effective number of migrants per island per generation, under these assumptions, is 10.1.

References

Bohonak, A.J. (2002). IBD (Isolation by Distance): A program for analyses of isolation by distance. *Journal of Heredity,* 93, 153-154.
Open access online at `http://jhered.oxfordjournals.org/content/93/2/153.full`

Excoffier, L., Laval, G., and Schneider, S. (2006). *Arlequin version 3.1: An Integrated Software Package for Population Genetics Data Analysis.* `http://cmpg.unibe.ch/software/arlequin3`

Fewster, R.M., Miller, S.D., and Ritchie, J. (2011). DNA profiling – a management tool for rat eradication. In: Veitch, C. R.; Clout, M. N. and Towns, D. R. (eds), *Island invasives: Eradication and management.* IUCN, Gland, Switzerland, p. 430-435.

Monmonier, M. (1973). Maximum-difference barriers: an alternative numerical regionalization method. *Geographical Analysis,* 3, 245-261.

Neigel, J.E. (2002). Is F_{ST} obsolete? *Conservation Genetics,* 3, 167-173.

Raymond, M. and Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity,* 86, 248-249.

Rousset, F. (2008). Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources,* 8, 103-106. `http://genepop.curtin.edu.au/`

Russell, J.C., and Fewster, R.M. (2009). Evaluation of the linkage disequilibrium method for estimating effective population size. In: Thomson, D. L., Cooch, E. G., Conroy, M. J. (eds), *Modeling Demographic Processes in Marked Populations.* Environmental and Ecological Statistics Series, Vol 3, Springer, Berlin, pp. 291-320.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution,* 38, 1358-1370.

Weir, B.S. (1996). *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.

Wright, S. (1978). *Evolution and the Genetics of Populations. Vol 4. Variability within and among natural populations.* University of Chicago Press, Chicago.