Random genetic drift

Lecture 5 ▪ Quantitative Genetics

Grum Gebreyesus ▪ ASAP-Bio Knowledge Hub

asap-bio.github.io/lec-qg-drift.html

Contents

1 Populations are finite

In deriving Hardy-Weinberg proportions we assumed the population was **infinite**, which is what guaranteed the constancy of allele frequencies. Real populations are finite. Once we admit that, allele frequencies stop being constant even in the complete absence of selection.

RECAP: RELATIONSHIP AND IBD

- Relationship can only be defined relative to a specific frame of reference, the base population.
- It rests on identity by descent: genes that are copies of the same gene in a past generation, or where one is a copy of the other, are IBD.
- θ_{ij} is the coefficient of coancestry, and $F_{ij} = 2\theta_{ij}$ the coefficient of relatedness.

2 What genetic drift is

RANDOM GENETIC DRIFT

Genetic drift describes the **random change in allele frequency** due to random sampling of alleles to form the offspring generation from the parent generation.

Where does the randomness come from?

- Any allele frequency that is neither 0 nor 1 can change at least a little, purely by chance in the sampling when one generation replaces another.
- A parent may, by chance at Mendelian segregation, transmit one allele more often than the other.
- Potential parents vary, partly by chance, in the number of offspring they actually have.

PROPERTIES OF THE PROCESS

- Drift is a **dispersive** process.
- Its effects can be quantified in **magnitude but not in direction**.
- Quantifying it requires very simplified assumptions.

Without selection, allele frequencies do not change *on average*. The frequency in the next generation is a random variable whose expectation is the current frequency, but whose variance is not zero. That distinction, no expected change but real variance, is the whole of drift.

3 The Fisher-Wright model

Take a population of N diploid individuals. They carry $2N$ alleles. Each offspring is formed by picking two parents at random *with replacement*, and picking one allele from each. There is no selection.

The allele frequency in the next generation is then a random variable. If among the parents there are n_t copies of A_1 out of $2N$ genes, each allele in the offspring is A_1 independently with probability $n_t/2N$, so the count follows a **binomial** distribution:

$$P(n_{t+1} = k \mid n_t, 2N) = \binom{2N}{k} \left(\frac{n_t}{2N}\right)^k \left(1 - \frac{n_t}{2N}\right)^{2N-k}$$

4 Fixation

In a finite population, this random walk has absorbing boundaries.

- Over time alleles get **fixed**; they reach 100%.
- Or they are **lost**; they reach a frequency of 0%.
- Without mutation or immigration, all loci will eventually become **monomorphic**.
- This happens relatively faster in a small population.

Selection may preserve polymorphism, for instance under overdominance, but drift alone will not.

5 The ideal population

The process of genetic drift can be quantified, but only if we assume simplified conditions. The reference is the **Fisher-Wright ideal population**:

THE IDEAL POPULATION

- N_e individuals, constant through generations, diploid and hermaphrodite.
- Random sampling of alleles for the next generation, *with replacement*.
- No selection, and mutation ignored.

Therefore all individuals have exactly the same chance of contributing, and mating between gametes is random.

Comparing a real population to this idealisation is what gives the effective population size its meaning.

6 Deriving the drift coefficient

We measure divergence of allele frequencies by F_{ST}, **Wright's drift coefficient**. It measures the diversification of sub-populations, and equally how far any one population has diverged from its origin. Let us derive how it accumulates.

From generation to generation

1. **Generation 0.** By definition of the base population, $F_{ST_0} = 0$.
2. **Generation 1.** Two randomly chosen genes are IBD only if they are copies of the *same* gene in generation 0. Since there are $2N_e$ genes to choose from, $F_{ST_1} = \frac{1}{2N_e}$.
3. **Generation 2.** Now there are *two* ways that two genes can be IBD:
   (1) they are copies of the *same* gene in generation 1, with probability $\frac{1}{2N_e}$; or
   (2) they are copies of *different* genes in generation 1 that were themselves IBD, with probability $\left(1 - \frac{1}{2N_e}\right) F_{ST_1}$.
   Adding them, $F_{ST_2} = \left(1 - \frac{1}{2N_e}\right) F_{ST_1} + \frac{1}{2N_e}$.
4. **The general recursion.** The same argument applies at every generation, so

$$F_{ST_{t+1}} = \left(1 - \frac{1}{2N_e}\right) F_{ST_t} + \frac{1}{2N_e}$$

Solving the recursion

1. Start from the recursion and subtract each side from one:

$$1 - F_{ST_{t+1}} = 1 - \frac{1}{2N_e} - \left(1 - \frac{1}{2N_e}\right) F_{ST_t}$$

2. Factorise the right-hand side:

$$1 - F_{ST_{t+1}} = \left(1 - \frac{1}{2N_e}\right)(1 - F_{ST_t})$$

   so the quantity $1 - F_{ST}$, the variation *remaining*, is multiplied by the same factor every generation.
3. Iterating from generation 0 gives

$$1 - F_{ST_t} = (1 - F_{ST_0})\left(1 - \frac{1}{2N_e}\right)^t$$

4. With $F_{ST_0} = 0$ this rearranges to the result we want:

$$\boxed{F_{ST_t} = 1 - \left(1 - \frac{1}{2N_e}\right)^t}$$

7 Effective population size

EFFECTIVE POPULATION SIZE

For a real population, the **effective population size** is the size of an *ideal* population that loses variation at the same rate as the real population does.

7.1 Rate of loss

The fraction of variance lost in one generation is

$$\Delta F_t = \frac{F_{ST_t} - F_{ST_{t-1}}}{1 - F_{ST_{t-1}}}$$

reading the numerator as the variance lost going from $t-1$ to t, and the denominator as the variance still left at $t-1$.

ΔF in terms of N_e

1. From the derivation above, $(1 - F_{ST_{t+1}}) = \left(1 - \frac{1}{2N_e}\right)(1 - F_{ST_t})$.
2. Substitute into the definition:

$$\Delta F_{ST_{t+1}} = \frac{(1 - F_{ST_t}) - (1 - F_{ST_{t+1}})}{1 - F_{ST_t}} = \frac{(1 - F_{ST_t}) - (1 - F_{ST_t})\left(1 - \frac{1}{2N_e}\right)}{1 - F_{ST_t}}$$

3. The factor $(1 - F_{ST_t})$ cancels, leaving $\Delta F = \frac{1}{2N_e}$, independent of t.

DRIFT EFFECTIVE POPULATION SIZE

$$\Delta F = \frac{1}{2N_e}$$

This is the definition of N_e used throughout. There are other, nearly equivalent, definitions.

8 What makes N_e smaller than N

The effective size of most populations is typically **much smaller** than the real number of individuals. Five mechanisms account for most of the gap.

8.1 Skewed sex ratio

$$N_e \approx \frac{4N_m N_f}{N_m + N_f}$$

with the upper bound $N_e < 4\min(N_m, N_f)$. This is very important in domestic animals, where a handful of AI sires may serve an enormous number of females, and in some natural populations.

8.2 Variance in offspring number

$$N_e \approx \frac{4N}{2 + \sigma^2}$$

When parents are chosen at random, offspring numbers are approximately Poisson, $\sigma^2 = 2$, and $N_e \approx N$. With selection among parents $\sigma^2 \gg 2$ and $N_e \ll N$. In many species only a tiny fraction of individuals reproduce at all, oak trees produce huge numbers of seeds of which very few establish.

8.3 Fluctuating population size

$$N_e \approx \frac{n}{\sum_{t=1}^{n} \frac{1}{N_t}}$$

This is a **harmonic** mean, so the *smallest* values dominate. Founder effects and bottlenecks belong in this category, and they are common in both natural and domesticated populations. One severe bottleneck permanently marks a population's N_e, however much it later recovers in census size.

8.4 Non-random mating

$$N_e \approx \frac{N}{1 + F}$$

where F is the deficiency of heterozygotes *due to non-random mating*. In domestic animals we deliberately mate non-relatives, so $F < 0$ and N_e is slightly increased. The magnitude is usually small.

8.5 Selection

Selection reduces the number of reproducing individuals, and typically means selecting *related* animals, so it reduces N_e. Quantitatively, with selection intensity i,

$$\Delta F \propto i^2, \qquad \Delta G \propto i$$

The asymmetry is important: gain rises *linearly* with intensity, whereas the loss of variation rises with its *square*. Increasing selection intensity therefore yields progressively less gain per unit of diversity lost.

9 Drift is not the same as inbreeding

A VERY IMPORTANT DISTINCTION

- Inbreeding due to **drift** and inbreeding due to **mating between relatives** are very different processes.
- Mating among relatives causes deviations from Hardy-Weinberg proportions, and can be *offset by a return to random mating*.
- Drift is a *loss of variation* through random change, and cannot be undone by changing the mating system.
- Both, however, cause an increase in the **average** F in the population.

Because both raise average F, they are readily confused when a single coefficient is read from a pedigree or a marker panel. The practical consequence is that a heterozygote deficiency caused by non-random mating disappears in one generation of random mating, whereas variation lost to drift is not recoverable.

10 F_{ST} as a variance ratio

When we worked with phenotypes we judged the importance of a factor by the variance it explains. The same logic applies to allele frequencies.

- Total variance available: $\bar{p}\,(1-\bar{p})$
- Variance explained by populations: $\mathrm{Var}(p_i)$

$$F_{ST} = \frac{\mathrm{Var}(p_i)}{\bar{p}\,(1-\bar{p})}$$

Worked example: three populations

Three populations have A at frequencies $p_1 = 0.2$, $p_2 = 0.4$, $p_3 = 0.6$.

- $\bar{p} = \sum_i p_i/3 = 0.4$, so $\bar{q} = 0.6$
- $\overline{p^2} = \sum_i p_i^2/3 = 0.1867$
- $\mathrm{Var}(p) = \overline{p^2} - (\bar{p})^2 = 0.1867 - 0.16 = 0.0267$

$$\hat{F}_{ST} = \frac{\mathrm{Var}(p)}{\bar{p}\,\bar{q}} = \frac{0.0267}{0.4 \times 0.6} = 0.111$$

So 11.1% of the variation is *between* populations.

TWO INTERPRETATIONS OF F_{ST}

- The random **differentiation of subpopulations**, if we have many.
- The random **change in the mean of one population**, over time.

These are the same quantity seen from two directions, which is why the same F_{ST} serves both the population-genetics literature on structure and the breeding literature on loss of variation.

11 Wright's F-statistics

Wright united the description of the two effects. An individual may be IBD at a locus because of genetic drift; if it is not, it may still be IBD because of non-random mating.

WRIGHT'S STATISTICS

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

The probability $(1 - F_{IT})$ of *not* being inbred is the probability of not being inbred due to non-random mating (F_{IS}) *and* not due to drift (F_{ST}).

The three coefficients are interpreted as follows:

- F_{IS} measures what is usually thought of as "inbreeding", that is, inbreeding due to mating between close relatives. It is typically positive, but may be slightly negative in managed populations where relatives are deliberately not mated.
- F_{ST} measures loss of variation *within* populations due to drift, and equally variation *between* populations due to drift. It is always ≥ 0.
- F_{IT} is the total, relative to the base population.

12 Generation time

Loss of variation, and therefore N_e, is typically measured *per generation*. That requires a definition of generation time, and it is arguable that per generation is not the most appropriate scale for comparing species or breeding schemes.

$$1 - \Delta F_{\text{gen}} = (1 - \Delta F_{\text{year}})^{T_g}$$

Two definitions are in use:

- **Forward looking:** the average age of members of a cohort when they have their offspring.
- **Backward looking:** the average age of the parents of members of a cohort.

This matters practically: a scheme that halves the generation interval doubles the rate of gain per year, but it also doubles the rate at which inbreeding accumulates per year.

13 Drift and quantitative traits

13.1 Inbreeding depression and heterosis

- Inbreeding leads to inbreeding depression. So does genetic drift.
- Selection against recessive deleterious alleles partially offsets some of this effect.

HETEROSIS

The superiority of hybrids over the average of their parents is due to the **cancellation of inbreeding depression** caused by drift. Hybrid vigour occurs when hybrids between two populations outperform the average of the parent populations.
This explains why crossing two long-separated inbred lines works so reliably: each line has drifted

to fixation for different deleterious recessives, and the cross restores heterozygosity at all of them.

13.2 Additive similarity does not care about the cause

Recall from the previous lecture that the covariance between the additive genetic effects of two individuals is $\mathrm{Cov}(a_i, a_j) = F_{ij}V_a$. Importantly, this says nothing about the *source* of F_{ij}, whether drift or mating between relatives. The algebra is identical; only the interpretation and the reversibility differ.

13.3 Change in the mean due to drift

- Over time F_{ST} builds up.
- In the limit $F_{ST} = 1$, all individuals in a population have the same genotype at every gene affecting the trait.
- The population mean is then the value of a *random* homozygous genotype from the base population.

The endpoint is random. Drift has no direction, so two replicate lines from the same base population fix at different means. This is exactly why drift must be taken into account when designing **selection experiments**, a divergence between selected and control lines is not evidence of response unless it exceeds what drift alone would produce.

14 Balancing gain against loss

SELECTION INTENSITY

The selection intensity i is the average phenotypic superiority of selected individuals over the population mean, in units of phenotypic standard deviation.

Stronger selection means more genetic gain, roughly $\Delta G \propto i$. But selection also reduces the number of reproducing individuals and concentrates them among relatives, so $\Delta F \propto i^2$.

BREEDING PLANNING

One of the major objectives of modern breeding planning is finding an **optimal balance** between ΔF and ΔG.

Because loss goes as i^2 while gain goes as i, the ratio $\Delta G/\Delta F$ falls as selection intensifies. There is no intensity that is simply "best"; the optimum depends on how the programme values long-term variance against short-term gain. This is the question that optimum contribution selection was invented to answer formally.

15 Summary

- Random change differentiates populations over time.
- The rate of change is measured by the effective population size.
- In most populations $N_e \ll N$.
- Heterosis occurs through the cancellation of inbreeding depression.
- Genetic drift also applies to differentiation in quantitative traits.

16 Exercises

Exercise 5.1

Estimating N_e from divergence. A large population is split into five separate parts of equal size. After 10 generations the allele frequencies at a locus are found to be 0.3, 0.4, 0.5, 0.6 and 0.7. Estimate the average N_e during the first 10 generations, assuming inbreeding in the base population is zero.

You will need $F_{ST} = \dfrac{\mathrm{Var}(p_i)}{\bar{p}(1-\bar{p})}$ and, after T generations, $1 - F_{ST_T} = \left(1 - \dfrac{1}{2N_e}\right)^T$.

Step 1, compute F_{ST}. $\bar{p} = (0.3 + 0.4 + 0.5 + 0.6 + 0.7)/5 = 0.5$.
$\mathrm{Var}(p_i) = \overline{p^2} - \bar{p}^2$. Here $\overline{p^2} = (0.09 + 0.16 + 0.25 + 0.36 + 0.49)/5 = 0.27$, so $\mathrm{Var}(p_i) = 0.27 - 0.25 = 0.02$.

$$F_{ST} = \frac{0.02}{0.5 \times 0.5} = 0.08$$

Step 2, solve for N_e. With $T = 10$,

$$1 - 0.08 = \left(1 - \frac{1}{2N_e}\right)^{10} \implies \left(1 - \frac{1}{2N_e}\right) = 0.92^{1/10} = 0.99169$$

$$\frac{1}{2N_e} = 0.00831 \implies N_e = \frac{1}{2 \times 0.00831} \approx 60$$

Each sub-population therefore behaved as an ideal population of approximately 60 individuals. This is an estimate of the *average* N_e over the 10 generations; where size fluctuates, the harmonic mean is dominated by the smallest generations.

Exercise 5.2

Simulating drift and fixation. Use the course R code to simulate allele-frequency change in a small population. The parameters are `tmax` (generations, large enough that most replicates fix), `nrep` (repetitions), `N` (population size) and `p0` (starting frequency).

```
 tmax <- 500; nrep <- 200; N <- 30; p0 <- 0.45
fixtime <- rep(NA, nrep)
plot(x=c(0,tmax), y=c(p0,p0), type="l", ylim=c(0,1),
xlab="Generation", ylab="Allele frequency", main=paste("Ne=",N))
lines(x=c(0,tmax), y=c(0,0)); lines(x=c(0,tmax), y=c(1,1))
```

1. Run it and describe the trajectories. What is the *fixation time*, and what is the *fixation probability*?
2. Re-run with $N = 30$ and then $N = 300$. How does population size change what you see?
3. Vary p_0. What determines the probability that the A allele is the one that fixes?

1. Each replicate performs a random walk until it hits 0 or 1, where it stays (an absorbing boundary). The *fixation time* is the generation at which polymorphism disappears; the *fixation probability* is the proportion of replicates in which the tracked allele reaches 1 rather than 0.
2. With $N = 30$ trajectories wander widely and most fix within a few hundred generations. With $N = 300$ the per-generation steps are much smaller (variance in p is of order $pq/2N$) and most replicates remain polymorphic for far longer. Mean time to fixation scales roughly with N_e.
3. Under pure drift the probability that an allele eventually fixes equals its *current frequency*, so with $p_0 = 0.45$ about 45% of replicates should fix for A. This follows from the fact that allele

frequency is a martingale under drift: there is no expected change, so the long-run probability of the absorbing state at 1 must equal the starting value.

Exercise 5.3

Comparing effective population sizes. A nucleus herd has 500 breeding females and 10 breeding sires. In a second scenario, the herd has 250 of each. Compute N_e for both, and comment on which design conserves more variation.

Using $N_e \approx \dfrac{4N_mN_f}{N_m + N_f}$.

Scenario 1: $N_e = \dfrac{4 \times 10 \times 500}{510} = \dfrac{20000}{510} \approx 39.$

Scenario 2: $N_e = \dfrac{4 \times 250 \times 250}{500} = \dfrac{250000}{500} = 500.$

Both herds contain 510 and 500 animals respectively, almost identical census sizes, yet the effective sizes differ by more than tenfold. The bound $N_e < 4\min(N_m, N_f) = 40$ makes the first scenario's ceiling explicit: with only 10 sires, no number of females can rescue N_e.

This is the principal practical implication for animal breeding: intense male selection, which artificial insemination makes possible and which maximises ΔG, also reduces N_e most rapidly.