Notes on Population Genetics

Graham Coop[1]

[1] Department of Evolution and Ecology & Center for Population Biology,

University of California, Davis.

To whom correspondence should be addressed: `gmcoop@ucdavis.edu`

Contents

1   Allele and Genotype frequencies

1.1   Allele frequencies

Consider a diploid autosomal locus segregating two alleles (1 and 2). Let's say that f_{11} is the frequency of 11 homozygotes and f_{12} is the frequency of 12 heterozygotes. The frequency of allele 1 in the population is

$$p = f_{11} + f_{12}/2 \tag{1}$$

note that this makes no assumption of Hardy-Weinberg [see below]. The frequency of the alternate allele (2) is then just $q = 1 - p$.

1.2   Hardy-Weinberg

Imagine a population mating at random with respect to our genotypes, i.e. no inbreeding, no population structure, no sex differences in allele frequencies.

    The frequency of allele 1 in the population at the time of reproduction is p. A 11 genotype is made by reaching out into our population and selecting two 1 alelle gametes to form a zygote. Therefore, the probability that our individual is a 11 homozygote is p^2. This probability is also the expected frequency of the 11 homozygote in the population.The expected frequency of our three genotypes is

f_{11}	f_{12}	f_{22}
p^2	$2pq$	q^2

Note that we only need to assume random mating with respect to our allele in order for these expected frequencies to hold, as long at p is the frequency of the 1 allele in the population at the time when gametes fuse.

1.3   Relatedness coefficients

We will define two alleles to be identical by descent if they are identical due to a common ancestor in the past few generations (for the moment ignoring the possibility of mutation). For example parent and child share exactly one allele identical by descent at a locus (assuming that the two parents of the child are randomly mated individuals from the population).

    A key quantity is the probability that our pair of individuals share 0, 1, or 2 alleles identical by descent. We denote these probabilities by r_0, r_1, and r_2 respectively. See Table 1 for some examples.

HapMap YRI (Africans)

HapMap CEU (Europeans)

Combined HapMap CEU + YRI (Europeans+Africans)

3
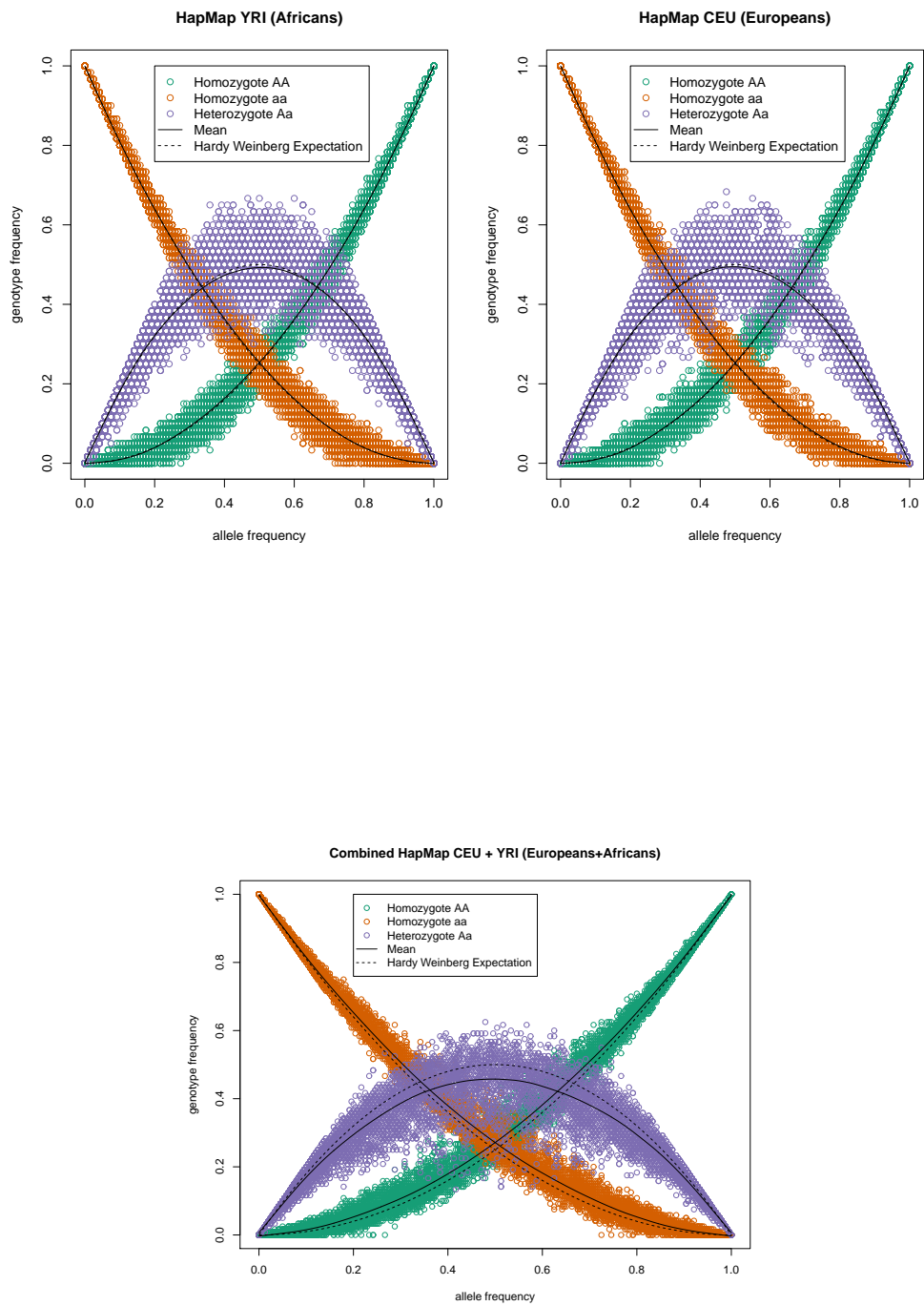
Relationship (i,j)*	r_0	r_1	r_2	F_{ij}
parent-child	0	1	0	1/4
full siblings	1/4	1/2	1/4	1/4
identical (monzygotic) twins	0	0	1	1/2
1st cousins	3/4	1/4	0	1/16

Table 1: Probability that two individuals of a given relationship share 0, 1, or 2 alleles identical by descent. * assuming this is the only relationship the pair of individuals share (above that expected from randomly sampling individuals from the population).

One summary of relatedness, which will be of help to us, is the probability that one allele picked at random from each of our two individuals is identical by descent. We call this quantity the coefficient of kinship of individual i and j, F_{ij}, and can we calculate it as

$$F_{ij} = 0 \times r_0 + \frac{1}{4}r_1 + \frac{1}{2}r_2 \tag{2}$$

This quantity will appear multiple times, in both our discussion of inbreeding and our discussion of the phenotypic resemblance between relatives.

1.4   Inbreeding

We can define an inbred individual as an individual whose parents are more closely related to each other than two random individuals drawn from some reference population.

When two related individuals produce an offspring, that individual can receive two alleles that are identical by descent, i.e. they can be homozygous by descent (sometimes termed autozygous), due to the fact that they have two copies of an allele through different paths through the pedigree. This increased likelihood of being homozygous relative to an outbred individual is the most obvious effect of inbreeding, and the one that will be of most interest to us as it underlies a lot of our ideas about inbreeding depression and population structure.

As the offspring receives a random allele from each parent (i and j) the probability that those two alleles are identical by descent is our kinship coefficient F_{ij} of the two parents (i.e. the quantity we defined in eqn. 2). This follows from the fact that our child's genotype is made by sampling an allele at random from each of our parents.

The only way the offspring can be heterozygous (A_1A_2) is if their two alleles at a locus, are not IBD (otherwise they'd necessarily be homozygous). Therefore, the probability that they are heterozygous is

$$(1 - F)2pq. \tag{3}$$

f_{11}	f_{12}	f_{22}
$(1-F)p^2 + Fp$	$(1-F)2pq$	$(1-F)q^2 + Fq$

Table 2: **Generalized Hardy Weinberg**

Our offspring can homozygous for the A_1 allele two different ways, they can have two non-IBD alleles which happen to be the A_1 allele , or their two alleles can be IBD, such that they inherited the A_1 by two different routes from the same ancestor. Thus the probability they are homozygous is

$$(1-F)p^2 + Fp \tag{4}$$

Therefore, our three genotype probabilities can be written as given in Table 2, generalizing our Hardy Weinberg proportions.

Note that the generalized Hardy Weinberg proportions completely specify our genotype probabilities, as there are two parameters (p and F) and two degrees of freedom (as our frequencies have to sum to one). Therefore, any combination of genotype frequencies at a biallelic site can be specified by a combination of p and F.

1.5   Calculating inbreeding coefficients from data

If the observed heterozygosity is f_{12} then an estimate of our inbreeding coefficient is

$$\hat{F} = 1 - \frac{f_{12}}{2pq} = \frac{2pq - f_{12}}{2pq} \tag{5}$$

where p is the frequency of the allele in our reference population. This can be rewritten in terms of the observed heterozygosity ($H_O = f_{12}$) and expected heterozygosity ($H_E = 2pq$)

$$\hat{F} = \frac{H_E - H_O}{H_E}. \tag{6}$$

If we have multiple loci we can replace H_O and H_E by their means over loci $\bar{H}_O$ and $\bar{H}_E$.

1.6   Summarizing Population structure

Our estimated inbreeding coefficient gives us a nice way to take a first look at population structure.

We defined inbreeding as having parents that are more closely related to each other than two random individuals drawn from some reference population. So the question naturally

arises, which reference population should we use? While I might not look inbred in comparison to allele frequencies in the UK (where I'm from), my parents certainly aren't too random individuals drawn from across the world-wide population. If we calculated F using allele frequencies within the UK, the inbreeding coefficient for me F would (hopefully) be close to zero, but would likely be larger if we used world-wide frequencies. That's because there's a somewhat lower level of heterozygosity within the UK than in the human population across the world as a whole.

Wright (1943, 1951) developed a set of 'F-statistics' (fixation statistics) that formalized these ideas about inbreeding. Wright defined F_{XY} as: the correlation between random gametes, drawn from the same X, relative to Y. We'll return to why F statistics are statements about correlations between alleles in just a moment. One commonly use F_{IS} for the inbreeding coefficient between an individual (I) and the subpopulation (S). Consider a single locus, where in a sub-population (S) a fraction $H_I = f_{12}$ of our individuals are heterozygotes. In this sub-population (S) the frequency of allele 1 is p_S and the expected heterozygosity is H_S. We will write F_{IS} as

$$F_{IS} = 1 - \frac{H_I}{H_S} = 1 - \frac{f_{12}}{2p_S q_S} \tag{7}$$

the direct analog of eqn. 5, which compares the observed heterozygosity to that expected under random mating within the sub-population. We could also compare our heterozygosity in individuals (H_I) to that expected in the total population (H_T). If the frequency of our allele in our total population is p_T, then we can write F_{IT} as

$$F_{IT} = 1 - \frac{H_I}{H_T} = 1 - \frac{f_{12}}{2p_T q_T} \tag{8}$$

which compares heterozygosity in individuals to that expected in the total population. As a simple extension of this we could imagine comparing the expected heterozygosity in the subpopulation (H_S) to that expected our total population (H_T), via F_{ST}

$$F_{ST} = 1 - \frac{H_S}{H_T} = 1 - \frac{2p_S q_S}{2p_T q_T} \tag{9}$$

If our total population contains our sub-population then, as we'll see below due to the Wahlund effect (to be added) $2p_S q_S \leq 2p_T q_T$ and so $F_{IS} \leq F_{IT}$ and $F_{ST} \geq 0$. We can relate our three F statistics together as

$$(1 - \hat{F_{IT}}) = \frac{H_I}{H_S}\frac{H_S}{H_T} = (1 - F_{IS})(1 - F_{ST}) \tag{10}$$

i.e. the reduction in heterozygosity in our individuals compared to that expected in the total population can be decomposed to the reduction in heterozygosity of individuals compared to the sub-population, and the reduction in heterozygosity from the total population to that

in the sub-population.

If we want a summary of population structure across multiple sub-populations we can average H_I and/or H_S across populations, and use a p_T calculated by averaging p_S across sub-populations. Furthermore, if we have multiple sites we can replace H_I, H_S, and H_T with their averages across loci (as above).

Lets now return to Wright's definition of the F_{XY} statistic as correlation between random gametes, drawn from the same X, relative to Y. With out loss of generality lets think about X as individuals and S as the sub-population. Rewriting F_{ST} in terms of our homozygote frequencies observed in individuals (f_{11} and f_{22}) we find

$$F_{IS} = \frac{2p_S q_S - f_{12}}{2p_S q_S} = \frac{f_{11} + f_{22} - p_S^2 - q_S^2}{2p_S q_S} \tag{11}$$

using the fact that $p^2 + 2pq + q^2 = 1$. The form of this (eqn. 11) is the covariance between pairs of alleles found in an individual, divided by the expected variance under binomial sampling. Thus F statistics can be understood as the correlation between alleles drawn from a population (or an individual) above that expected by chance (i.e. drawing alleles sampled at random from some broader population).

We can also see F statistics as proportions of variance explained by substructure. To see this lets think about F_{ST} averaged over K subpopulations, whose frequencies are $p_1, \cdots, p_K$. The frequency in the total population is $p_T = \bar{p} = \frac{1}{K} \sum_{i=1}^{K} p_i$. Then we can write

$$F_{ST} = \frac{2\bar{p}\bar{q} - \frac{1}{K}\sum_{i=1}^{K} 2p_i q_i}{2\bar{p}\bar{q}} = \frac{\left(\frac{1}{K}\sum_{i=1}^{K} p_i^2 + \frac{1}{K}\sum_{i=1}^{K} q_i^2\right) - \bar{p}^2 - \bar{q}^2}{2\bar{p}\bar{q}} = \frac{Var(p_i)}{Var(\bar{p})} \tag{12}$$

i.e. F_{ST} is the proportion of the variance explained by the subpopulation labels.

2 The phenotypic resemblance between relatives

added μ, mean to traits below

We can use our understanding of the sharing of alleles between relatives to understand the phenotypic resemblance between relatives in quantitative phenotypes. We can then use this to understand the evolutionary change in quantitative phenotypes in response to selection.

Let's imagine that the genetic component of the variation in our trait is controlled by L autosomal loci that act in an additive manner. The frequency of allele 1 at locus l is p_l, with each copy of allele 1 at this locus increasing your trait value by a_l. The phenotype of an individual, let's call her i, is X_i. Her genotype at SNP l, is $G_{i,l}$. Here $G_{i,l} = 0$, 1, or 2 represents the number of copies of allele 1 she has at this SNP. Her expected phenotype, given her genotype, is then

$$X_{A,i} = \mu + \mathbb{E}(X_i|G_{i,1}, \cdots, G_{i,L}) = \mu + \sum_{l=1}^{L} G_{i,l} a_l \tag{13}$$

where μ is the mean phenotype in our population. Now in reality the genetic phenotype is a function of the expression of those alleles in a particular environment. Therefore, we can think of this expected phenotype as being an average across a set of environments that occur in the population.

NEED to resolve μ in above equation

When we measure our individual's phenotype we see

$$X_i = \mu + X_{A,i} + X_{E,i} \tag{14}$$

where X_E is the deviation from the mean phenotype due to the environment. This X_E included the systematic effects of the environment our individual finds herself in and all of the noise during development, growth, and the various random insults that life throws at our individual. If a reasonable number of loci contribute to variation in our trait then we can approximate the distribution of $X_{A,i}$ by a normal distribution due to the central limit theorem (see R exercise). Thus if we can approximate the distribution of the effect of environmental variation on our trait ($X_{E,i}$) also by a normal distribution, which is reasonable as there are many small environmental effects, then the distribution of phenotypes within the population (X_i) will be normally distributed (see Figure 1).

Note that as this is an additive model we can decompose eqn. 14 into the effects of the two alleles at each locus, in particular we can rewrite it as

$$X_i = \mu + X_{iM} + X_{iP} + X_{iE} \tag{15}$$

where X_{iM} and X_{iP} are the contribution to the phenotype of the allele that our individual received from her mother (maternal alleles) and father (paternal alleles) respectively. This
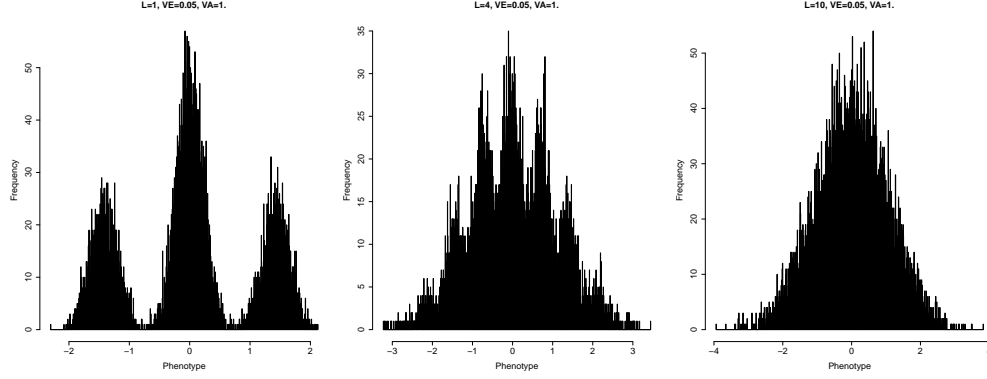
Figure 1: The convergence of the phenotypic distribution to a normal distribution.

will come in handy in just a moment when we start thinking about the phenotype covariance of relatives.

Now obviously this model seems silly at first sight as alleles don't only act in an additive manner, as they interact with alleles at the same loci (dominance) and at different loci (epistasis). Later we'll relax this assumption, however, we'll find that if we are interested in evolutionary change over short time-scales it is actually only the "additive component" of genetic variation that will (usually) concern us. We will define this more formally later on, but for the moment we can offer the intuition that parents only get to pass on a single allele at each locus on to the next generation. As such, it is the effect of these transmitted alleles, averaged over possible matings, that is an individual's average contribution to the next generation (i.e. the additive effect of the alleles that their genotype consists of).

2.0.1 Additive genetic variance and heritability

As we are talking about an additive genetic model we'll talk about the additive genetic variance (V_A), the variance due to the additive effects of segregating genetic variation. This is a subset of the total genetic variance if we allow for non-additive effects.

The variance of our phenotype across individuals (V) can write this as

$$V = Var(X_A) + Var(X_E) = V_A + V_E \tag{16}$$

in writing this we are assuming that there is no covariance between $X_{G,i}$ and $X_{E,i}$ i.e. there is no covariance between genotype and environment.

Our additive genetic variance can be written as

$$V_A = \sum_{l=1}^{L} Var(G_{i,l}a_l) \tag{17}$$

9

where $Var(G_{i,l}a_l)$ is the contribution to the additive variance among individuals of the l locus. Assuming random mating we can write our additive genetic variance as

$$V_A = \sum_{l=1}^{L} a_l^2 2p_l(1-p_l) \tag{18}$$

where the $2p_l(1-p_l)$ term follows the binomial sampling of two alleles per individual at each locus.

The narrow sense heritability   We would like a way to think about what proportion of the variation in our phenotype across individuals is due to genetic differences as opposed to environmental differences. Such a quantity will be key in helping us think about the evolution of phenotypes. For example, if variation in our phenotype had no genetic basis then no matter how much selection changes the mean phenotype within a generation the trait will not change over generations.

We'll call the proportion of the variance that is genetic the heritability, and denote it by h^2. We can then write this as

$$h^2 = \frac{Var(X_A)}{V} = \frac{V_A}{V} \tag{19}$$

remember that we thinking about a trait where all of the alleles act in a perfectly additive manner. In this case our heritability h^2 is referred to as the narrow sense heritability, the proportion of the variance explained by the additive effect of our loci. When we allow dominance and epistasis into our model we'll also have to define the broad sense heritability (the total proportion of the phenotypic variance attributable to genetic variation).

The narrow sense heritability of a trait is a useful quantity, indeed we'll see shortly that it is exactly what we need to understand the evolutionary response to selection on a quantitative phenotype. We can calculate the narrow sense heritability by using the resemblance between relatives. For example, if our phenotype was totally environmental we should not expect relatives to resemble each other any more than random individuals drawn from the population. Now the obvious caveat here is that relatives also share an environment, so may resemble each other due to shared environmental effects.

2.0.2   The covariance between relatives

So we'll go ahead and calculate the covariance in phenotype between two individuals (1 and 2) who have a phenotype X_1 and X_2 respectively.

$$Cov(X_1, X_2) = Cov\left((X_{1M} + X_{1P} + X_{1E}), ((X_{2M} + X_{2P} + X_{2E}))\right) \tag{20}$$

We can expand this out in terms of the covariance between the various components in these sums.

To make our task easier we (and most analyses) will assume two things

1. that we can ignore the covariance of the environments between individuals (i.e. $Cov(X_{1E}, X_{2E}) = 0$)

2. that we can ignore the covariance between the environment variation experience by an individual and the genetic variation in another individual (i.e. $Cov(X_{1E}, (X_{2M} + X_{2P})) = 0$).

The failure of these assumptions to hold can severely undermine our estimates of heritability, but we'll return to that later. Moving forward with these assumptions, we can write our phenotypic covariance between our pair of individuals as

$$Cov(X_1, X_2) = Cov((X_{1M}, X_{2M}) + Cov(X_{1M}, X_{2P}) + Cov(X_{1P}, X_{2M}) + Cov(X_{1P}, X_{2P}) \quad (21)$$

This is saying that under our simple additive model we can see the covariance in phenotypes between individuals as the covariance between the allelic effects in our individuals. We can use our results about the sharing of alleles between relatives to obtain these terms. But before we write down the general case lets quickly work through some examples.

The covariance between Identical Twins Lets first consider the case of a pair of identical twins from two unrelated parents. Our pair of twins share their maternal and paternal allele identical by descent ($X_{1M} = X_{2M}$ and $X_{1P} = X_{2P}$). As their maternal and paternal alleles are not correlated draws from the population, i.e. have no probability of being IBD as we've said the parents are unrelated, the covariance between their effects on the phenotype is zero (i.e. $Cov(X_{1P}, X_{2M}) = Cov(X_{1M}, X_{2P}) = 0$). In that case eqn. 21 is

$$Cov(X_1, X_2) = Cov((X_{1M}, X_{2M}) + Cov(X_{1P}, X_{2P}) = 2Var(X_{1M}) = V_A \quad (22)$$

Now in general identical twins are not going to be super helpful for us in estimating h^2 as under models with non-additive effects identical twins have higher covariance than we'd expect as they resemble each other also because of the dominance effects as they don't just share alleles they share their entire genotype.

The covariance in phenotype between mother and child . If the mother and father are unrelated individuals (i.e. are two random draws from the population) then the mother and a child share one allele IBD at each locus (i.e. $r_1 = 1$ and $r_0 = r_2 = 0$). Half the time our mother transmits her paternal allele to the child, in which case $X_{P1} = X_{M2}$ and so $Cov(X_{P1}, X_{M2}) = Var(X_{P1})$ and all the other covariances in eqn. 21 zero, and half the time she transmits her maternal allele to the child $Cov(X_{M1}, X_{M2}) = Var(X_{M1})$ and all the other terms zero. By this argument $Cov(X_1, X_2) = \frac{1}{2}Var(X_{M1}) + \frac{1}{2}Var(X_{P1}) = \frac{1}{2}V_A$.

The covariance between general pairs of relatives under an additive model The two examples make clear that to understand the covariance between phenotypes of relatives we simply need to think about the alleles they share IBD. Consider a pair of relatives (1 and 1) with a probability r_0, r_1, and r_2 of sharing zero, one, or two alleles IBD respectively. When they share zero alleles $Cov((X_{1M} + X_{1P}), (X_{2M} + X_{2P})) = 0$, when they share one allele $Cov((X_{1M} + X_{1P}), (X_{2M} + X_{2P})) = Var(X_{1M}) = \frac{1}{2}V_A$, and when they share two alleles $Cov((X_{1M} + X_{1P}), (X_{2M} + X_{2P})) = V_A$. Therefore, the general covariance between two relatives is

$$Cov(X_1, X_2) = r_0 \times 0 + r_1 \frac{1}{2}V_A + r_2 V_A = 2F_{1,2}V_A \tag{23}$$

I seem to have been missing a factor of two in the above equations. I apologize for the confusion caused. So under a simple additive model of the genetic basis of a phenotype to measure the narrow sense heritability we need to measure the covariance between a set of pairs of relatives (assuming that we can remove the effect of shared environmental noise). From the covariance between relatives we can calculate V_A, we can then divide this by the total phenotypic variance to get h^2.

Another way that we can estimate the narrow sense heritability is through the regression of child's phenotype on the parental mid-point phenotype. The parental mid-point phenotype is simple the average of the mum and dad's phenotype. Denoting the child's phenotype by X_{kid} and mid-point phenotype by X_{mid} so that if we take the regression $X_{kid} \sim X_{mid}$ this regression has slope $\beta = Cov(X_{kid}, X_{mid})/Var(X_{mid})$. The covariance of $Cov(X_{kid}, X_{mid}) = \frac{1}{2}V_A$, and $Var(X_{mid}) = \frac{1}{2}V$ as by taking the average of the parents we have halved the variance, such that the slope of the regression is

$$\beta = \frac{Cov(X_{kid}, X_{mid})}{Var(X_{mid})} = \frac{V_A}{V} = h^2 \tag{24}$$

i.e. the regression of the child's phenotype on the parental midpoint phenotype is an estimate of the narrow sense heritability. This is a common way to estimate heritability, although it doesn't bypass the need to control for environmental correlations between relatives.

Our regression allows us to attempt to predict the phenotype of the child given the child; how well we can do this depends on the slope. If the slope is close to zero then the parental phenotypes hold no information about the phenotype of the child, while if the slope is close to one then the parental mid-point is a good guess at the child's phenotype.

More formally the expected phenotype of the child given the parental phenotypes is

$$\mathbb{E}(X_{kid}|X_{mum}, X_{dad}) = \mu + \beta(X_{mid} - \mu) = \mu + h^2(X_{mid} - \mu) \tag{25}$$

this follows from the definition of linear regression. So to find the child's predicted phenotype we simply take the mean phenotype and add on the difference between our parental mid-point multiplied by our narrow sense heritability.
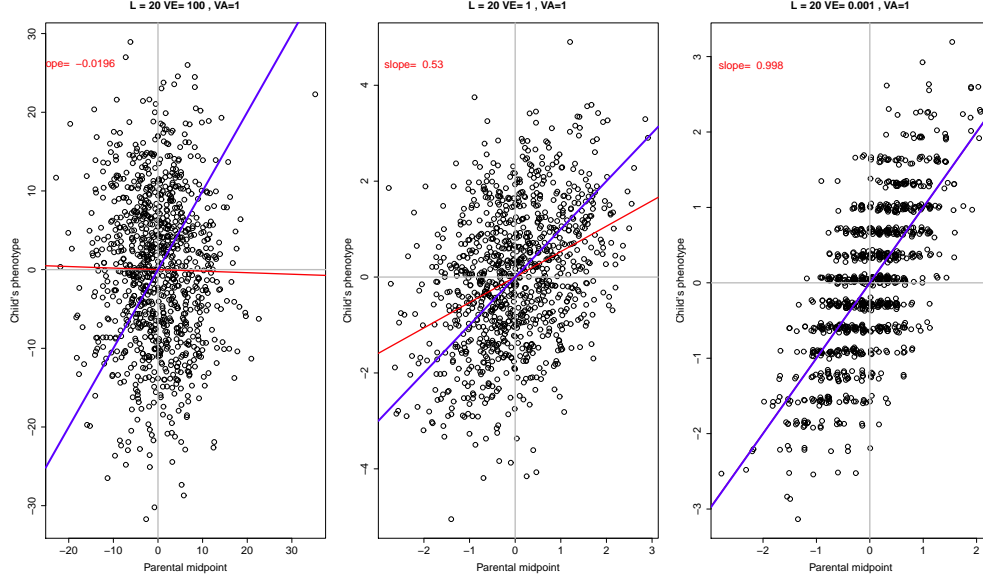
Figure 2: Regression of parental mid-point phenotype on child's phenotype.

Estimating additive genetic variance across a variety of different relationships.
New material In many natural populations we may have access to individuals of a range
of different relationships to each other (through monitoring of the paternity of individuals),
but relatively few individuals of a given relationship (e.g. sibs). We can try and use this
information as fully as possible in a mixed model framework. Considering equation 14 we
can write an individual's phenotype X_i as

$$X_i = \mu + X_{A,i} + e_i \tag{26}$$

where $e_i \sim N(0, V_E)$ and $X_{A,i}$ is normally distributed across individuals with covariance
matrix $V_A A$ where the the entries for an pair of individuals i and i are $A_{ij} = 2F_{i,j}$ and $A_{ii}1$.
Given the matrix A we can estimate V_A. We can also add fixed effects into this model to
account for generation effects, additional mixed effects could also be included to account for
shared environments between particular individuals (e.g. a shared nest). This is sometimes
called the "animal model", and also goes by the name of variance components analysis.

Multiple traits. *New material* Traits often covary with each other, due to both environ-
mentally induced effects (e.g. due to the effects of diet on multiple traits) and due to the
expression of underlying genetic covariance between traits. In turn this genetic covariance
can reflect pleiotropy, a mechanistic effect of an allele on multiple traits (e.g. variants that
effect skin pigmentation often effect hair color) or the genetic linkage of loci independently
affecting multiple traits. If we are interested in evolution over short time-scales we can
(often) ignore the genetic basis of this correlation.

Consider two traits $X_{1,i}$ and $X_{2,i}$ in an indivdual i, these could be say the indivdual's leg length and nose length. As before we can write these as

$$\begin{aligned} X_{1,i} &= \mu_1 + X_{1,A,i} + X_{1,E,i} \\ X_{2,i} &= \mu_2 + X_{2,A,i} + X_{2,E,i} \end{aligned} \tag{27}$$

As before we can talk about the total phenotypic variance (V_1, V_2), environmental variance $(V_{1,E}$ and $V_{2,E})$, and the additive variance and in trait one and two and $(V_{1,A}, V_{2,A})$. But now we also have to consider the total covariance $V_{1,2} = Cov(X_1, X_2)$, the environmentally induced covariance between the traits $(V_{E,1,2} = Cov(X_{1,E}, X_{2,E}))$ and the additive genetic covariance $(V_{A,1,2} = Cov(X_{1,A}, X_{2,A}))$ between trait one and two.

We can store these values in a matrices

$$\mathbf{V} = \begin{pmatrix} V_1 & V_{1,2} \\ V_{1,2} & V_2 \end{pmatrix} \tag{28}$$

and

$$\mathbf{G} = \begin{pmatrix} V_{1,A} & V_{A,1,2} \\ V_{A,1,2} & V_{2,A} \end{pmatrix} \tag{29}$$

we can generalize this to an abitrary number of traits.

We can estimate these quantities, in a similar way to before, by studying the covariance in different traits between relatives:

$$Cov(X_{1,i}, X_{2,j}) = 2F_{i,j} V_{A,1,2} \tag{30}$$

2.0.3 The response to selection

Evolution by natural selection requires:

1. Variation in a phenotype

2. That survival is non-random with respect to this phenotypic variation.

3. That this variation is heritable.

Points 1 and 2 encapsulate our idea of Natural Selection, but evolution by natural selection will only occur if the 3rd condition is met. It is the heritable nature of variation that couples change within a generation due to natural selection, to change across generations (evolutionary change).

Lets start by thinking about the change within a generation due to directional selection, where selection acts to change the mean phenotype within a generation. For example, a

14

decrease in mean height within a generation, due to taller organisms having a lower chance of surviving to reproduction than shorter organisms. Specifically, we'll denote our mean phenotype at reproduction by μ_S, i.e. after selection has acted, and our mean phenotype before selection acts by μ_{BS}. This second quantity may be hard to measure, as obviously selection acts throughout the life-cycle, so it might be easier to think of this as the mean phenotype if selection hadn't acted. So the mean phenotype changes within a generation is $\mu_S - \mu_{BS} = S$.

We are interested in predicting the distribution of phenotypes in next generation, in particular we are interested in the mean phenotype in the next generation to understand how directional selection has contributed to evolutionary change. We'll denote the mean phenotype in offspring, i.e. the mean phenotype in the next generation before selection acts, as μ_{NG}. The change across generations we'll call the response to selection R and put this equal to $\mu_{NG} - \mu_{BS}$.

The mean phenotype in the next generation is

$$\mu_{NG} = \mathbb{E}\left(\mathbb{E}(X_{kid}|X_{mum}, X_{dad})\right) \tag{31}$$

where the outer expectation is over the randomly mating of individuals who survive to reproduce. We can use eqn. 25 to obtain an expression for this

$$\mu_{NG} = \mu_{BS} + \beta(\mathbb{E}(X_{mid}) - \mu_{BS}) \tag{32}$$

so to obtain μ_{NG} we need to compute $\mathbb{E}(X_{mid})$ the expected mid-point phenotype of pairs of individuals who survive to reproduce. Well this is just the expected phenotype in the individuals who survived to reproduce (μ_S), so

$$\mu_{NG} = \mu_{BS} + h^2(\mu_S - \mu_{BS}) \tag{33}$$

So we can write our response to selection as

$$R = \mu_{NG} - \mu_{BS} = h^2(\mu_S - \mu_{BS}) = h^2 S \tag{34}$$

So our response to selection is proportional to our selection differential, and the constant of proportionality is the narrow sense heritability. This equation is sometimes termed the Breeders equation. It is a statement that the evolutionary change across generations (R) is proportional to the change caused by directional selection within a generation, and the strength of this relationship is determined by the narrow sense heritability.

Using the fact that $h^2 = V_A/V$ we can rewrite this in a different form as

$$R = V_A \frac{S}{V} \tag{35}$$

Phenotype distribution before selection, Mean=0, VA=1, VE=1, Taking top 10%

Phenotype distribution after selection, parental mean= 2.474

Phenotype distribution in the children Mean in children = 1.208
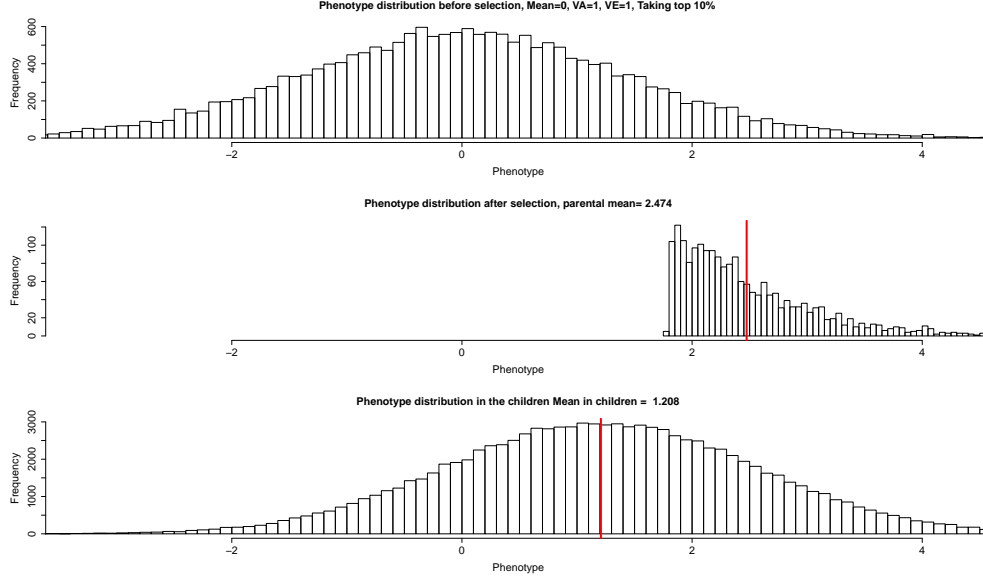
i.e. our response to selection is the additive genetic variance of our trait (V_A) multiplied by the change within a generation as a fraction of the total phenotypic variance (S/V, *sometimes called the the selection gradient β*).

A change in mean phenotype within a generation occurs because of the differential fitness of our organisms. To think more carefully about this change within a generation lets think about a simple fitness model where our phenotype affects the viability of our organisms (i.e. the probability they survive to reproduce). The probability that an individual has a phenotype X before selection is $p(X)$, so that the mean phenotype before selection is

$$\mu_{BS} = \mathbb{E}[X] = \int_{-\infty}^{\infty} x p(x) dx \tag{36}$$

The probability that an organism with a phenotype X survives to reproduce is $w(X)$, and we'll think about this as the fitness of our organism. The probability distribution of phenotypes in those who do reproduce is

$$\mathbb{P}(X|\text{survive}) = \frac{p(x)w(x)}{\int_{-\infty}^{\infty} p(x)w(x)dx}. \tag{37}$$

where the denominator is a normalization constant which ensures that our phenotypic distribution integrates to one. The denominator also has the interpretation of being the mean fitness of the population, which we'll call $\overline{w}$, i.e.

$$\overline{w} = \int_{-\infty}^{\infty} p(x)w(x)dx. \tag{38}$$

16

Therefore, we can write the mean phenotype in those who survive to reproduce as

$$\mu_S = \frac{1}{\overline{w}} \int_{-\infty}^{\infty} xp(x)w(x)dx \tag{39}$$

If we mean center our population, i.e. set the phenotype before selection to zero, then

$$S = \frac{1}{\overline{w}} \int_{-\infty}^{\infty} xp(x)w(x)dx \tag{40}$$

if $\mu_S = 0$. Inspecting this more closely we can see that S has the form of a covariance between our phenotype X and our fitness $w(X)$ ($Cov(X, w(X))$). Thus our change in mean phenotype is directly a measure of the covariance of our phenotype and our fitness. Rewriting our breeder's equation using this observation we see

$$R = \frac{V_A}{V} Cov(X, w(X)) \tag{41}$$

we see that the response to selection is due to the fact that our fitness (viability) of our organisms/parents covaries with our phenotype, and that our child's phenotype is correlated with the parent phenotype.

The response of multiple traits to selection, the multivariate breeder's equation.
New material We can generalize these results for multiple traits, to ask how selection on multiple phenotypes plays out over short time intervals. We'll write our change in the mean our multiple phenotypes within a generation as the vector $\mathbf{S}$ and our response across multiple generations as the vector $\mathbf{R}$. These two quantities are related by

$$\mathbf{R} = \mathbf{G}\mathbf{V}^{-1}\mathbf{S} = \mathbf{G}\beta \tag{42}$$

where $\mathbf{V}$ and $\mathbf{G}$ are our matrices of the variance-covariance of phenotypes and additive genetic values (eqn. (29) (28)) and β is a vector of selection gradients (i.e. the change within a generation as a fraction of the total phenotypic variance). To make this a bit more intuitive, consider two traits we are writing

$$\begin{aligned} R_1 &= V_{A,1}\beta_1 + V_{A,1,2}\beta_2 \\ R_2 &= V_{A,2}\beta_2 + V_{A,1,2}\beta_1 \end{aligned}$$

$$\tag{43}$$

where the 1 and 2 index our two different traits. This is a statement that our response in any one phenotype is modified by selection on other traits that covary with that trait. This offers a good way to think about how genetic trade offs play out over evolution over short time-scales.

Breeding values *Here be dragons, this section has not be worked through* Before we turn to our discussion of the response to selection it will help to define one more quantity, the breeding value. The breeding value of an individual is the mean phenotype of the offspring of this individual averaged over all possible matings in the population. This is X_{Ai}.

For an individual i we can write the the breeding value of an individual G_X as

$$G_i = \alpha_G + \beta_{GP} X_i + \epsilon_G \tag{44}$$

where $\beta_{GX} Cov(X, G)/Var(X) = h^2$

The breeder's equation is in fact a statement of the change in the mean breeding value of the population. To see this consider the fact that we can write the change in mean breeding value within a generation due to selection as

$$R = \mathbb{E}[G_i] = \beta_{GP} \bar{X} - \bar{P_X} \tag{45}$$

We can also write the breeder's equation as

$$R = \delta \bar{G} = Cov(G, W) \tag{46}$$

i.e. the mean change in phenotypes over the generations is due to the fact that our breeding values covaries with fitness this is the Price equation.

If ϵ_G and G are uncorrelated then

$$\delta \bar{G} = Cov(G, W) = \beta_{GP} Cov(P, W) = h^2 S \tag{47}$$

we can also write

$$W = \alpha_W + \beta_{WP} P + \epsilon_W \tag{48}$$

where $\beta_{WP} = Cov(W, P)/Var(P)$. If ϵ_W and G are uncorrelated

$$\delta \bar{G} = Cov(G, W) = \beta_{WP} Cov(G, P) \tag{49}$$

2.0.4 Hamiliton's Rule and the evolution of altruistic and selfish behaviours

Here be dragons, this section has not be worked through We can use our quantitative genetics framework to gain some simple intuition for when altruistic behaviours should evolve through kin selection. To do this we can follow Queller's (1992) treatment, to rederive Hamiliton's rule in a quantitative genetics framework (Hamilition original work did this in a population genetics framework).

So lets imagine that individuals interact in pairs, with our focal individual i being paired with an individual j. Imagine that individuals have two possible phenotypes $X = 1$ or 0, corresponding to providing or withholding some small act of 'Altruism' (we could just as easily flip these labels and call them a unselfish act and a selfish act respectively). Providing the altruistic act has a cost C to the fitness of our individual, with failing to provide this act has no cost. Receiving this altruistic act confers a fitness benefit B (not receiving it has

no benefit). Hamiliton's Rule is that such a trait will spread through the population if the average coefficient of relatedness between two individuals is r, i.e. the probability that a gene picked randomly from each at the same locus is identical by descent, such that $C < rB$. The intuition here is that while our focal individual i is losing out of some reproductive output, the alleles underlying the behavior are still spreading as this cost is outweighed by the transmission of these alleles through a related individual (note that this means that the allele is not acting in an self sacrificing manner, even though individuals may as a result).

We will assume that our focal i individual's relative fitness can be written as

$$W(X_i, X_j) = W_0 + W_i + W_j \tag{50}$$

where W_i is the contribution of the fitness of the individual i due to this phenotype, and W_j is the contribution to our individual i's fitness due to the j's behaviour (i.e. phenotype). With the benefit B and cost C our $W(X_i, X_j)$ are depicted in Figure XXX. We can write our fitness component for our individual's fitness as

$$W_i = \alpha_i + \beta_{W_i P_i} X_i + \epsilon_{W_i} \tag{51}$$

and

$$W_j = \alpha_j + \beta_{W_j P_j} P_j + \epsilon_{W_j} \tag{52}$$

our slopes $\beta_{W_i P_i} = Cov(W_i, P_i)/Var(P_i)$ and $\beta_{W_j P_j} = Cov(W_j, P_j)/Var(P_j)$ i.e. the regression our our individual is fitness components on his and his neighbour's phenotype. From Figure XXX these slopes are just our cost and benefit.

Using the price equation, assuming that the residuals are not correlated with our genotype we arrive at

$$\delta \bar{G}_X = \beta_{W_i P_i} Cov(G_i, P_i) + \beta_{W_j P_j} Cov(G_i, P_j) \tag{53}$$

so our altruistic trait will increase in frequency in the population only if $\delta \bar{G}_X > 0$, i.e. that

$$0 < \beta_{W_i P_i} + \beta_{W_j P_j} \frac{Cov(G_i, P_j)}{Cov(G_i, P_i)} \tag{54}$$

substituting in our slopes we arrive at

$$C < B \frac{Cov(G_i, P_j)}{Cov(G_i, P_i)} \tag{55}$$

i.e. our altruistic trait will increase if the cost is outweighed by the benefit weighted by a factor $\frac{Cov(G_i, P_j)}{Cov(G_i, P_i)}$. The term $\frac{Cov(G_i, P_j)}{Cov(G_i, P_i)}$ is the ratio of the ratio of covariance between our focal individual's genotype and their phenotype and that of their interacting neighbour.

3   Correlations between loci, linkage disequilibrium, and recombination.

Up to now we've been interested in correlations between alleles at the same locus, e.g. correlations within individuals (inbreeding) or between individuals (relatedness). We turn our attention now to think about correlations between alleles at different loci. To start to understand correlations between loci we need to first understand a bit about recombination.

Recombination   Lets consider an individual heterozygous for a AB and ab haplotype. If no recombination occurs between our two loci in this individual, then these two haplotypes will be transmitted intact to the next generation. While if a recombination (or more generally an odd number of recombinations) occurs between our two loci on the haplotype transmitted to the child then $\frac{1}{2}$ the time the child receives a Ab haplotype and $\frac{1}{2}$ the time the child receives a aB haplotype. So recombination is breaking up the association between loci. We'll define the recombination fraction (r) to be the probability of an odd number of recombinations between our loci. In practice we'll often be interested in relatively short regions where recombination is relatively rare, and so we might think that $r = r_{BP}L \ll 1$, where r_{BP} is the average recombination rate per base pair (typically $\sim 10^{-8}$) and L is the number of base pairs separating our two loci.

Linkage disequilibrium   The (horrible) phrase linkage disequilibrium (LD) refers to the statistical non-independence (i.e. a correlation) of alleles at different loci. Our two loci, which segregate alleles A/a and B/b, have allele frequencies of p_A and p_B respectively. The frequency of the two locus haplotype is p_{AB}, and likewise for our other three combinations. If our loci were statistically independent then $p_{AB} = p_A p_B$, otherwise $p_{AB} \neq p_A p_B$ We can define a covariance between the A and B alleles at our two loci as

$$D_{AB} = p_{AB} - p_A p_B \tag{56}$$

and likewise for our other combinations at our two loci (D_{Ab}, D_{aB}, D_{ab}). These D statistics are all closely related to each other as $D_{AB} = -D_{Ab}$ and so on. Thus we only need to specify one D_{AB} to know them all, so we'll drop the subscript and just refer to D. Also a handy result is that we can rewrite our haplotype frequency p_{AB} as

$$p_{AB} = p_A p_B + D. \tag{57}$$

If $D = 0$ we'll say the two loci are in linkage equilibrium, while if $D > 0$ or $D < 0$ we'll say that the loci are in linkage disequilibrium (we'll perhaps want to test whether D is statistically different from 0 before making this choice). You should be careful to keep the concepts of linkage and linkage disequilibrium separate in your mind. Genetic linkage refers to the linkage of multiple loci due to the fact that they are transmitted through meiosis together

(most often because the loci are on the same chromosome). Linkage disequilibrium merely refers to the correlation between the alleles at different loci, this may in part be due to the genetic linkage of these loci but does not necessarily imply this (e.g. genetically unlinked loci can be in LD due to population structure).

Another common statistic for summarizing LD is r^2 which we write as

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \tag{58}$$

as D is a covariance, and $p_A(1-p_A)$ is the variance of an allele drawn at random from locus A, r^2 is the squared correlation coefficient.

Question. You genotype 2 bi-allelic loci (A & B) segregating in two mouse subspecies (1 & 2) which mate randomly among themselves, but have not historically interbreed since they speciated. On the basis of previous work you estimate that the two loci are separated by a recombination fraction of 0.1. The frequencies of haplotypes in each population are:

Pop	p_{AB}	p_{Ab}	p_{aB}	p_{ab}
1	.02	.18	.08	.72
2	.72	.18	.08	.02

A) How much LD is there within populations, i.e. estimate D?

B) If we mixed the two populations together in equal proportions what value would D take before any mating has had the chance to occur?

The decay of LD due to recombination We will now examine what happens to LD over the generations if we only allow recombination to occur in a very large population (i.e. no genetic drift, i.e. the frequencies of our loci follow their expectations). To do so consider the frequency of our AB haplotype in the next generation p'_{AB}. We lose a fraction r of our AB haplotypes to recombination ripping our alleles apart but gain a fraction rp_Ap_B per generation from other haplotypes recombining together to form AB haplotypes. Thus in the next generation

$$p'_{AB} = (1-r)p_{AB} + rp_Ap_B \tag{59}$$

this last term here is $r(p_{AB} + p_{Ab})(p_{AB} + p_{aB})$, which multiplying this out is the probability of recombination in the different diploid genotypes that could generate a p_{AB} haplotype.

We can then write the change in the frequency of the p_{AB} haplotype as

$$\Delta p_{AB} = p'_{AB} - p_{AB} = -rp_{AB} + rp_Ap_B = -rD \tag{60}$$

so recombination will cause a decrease in the frequency of p_{AB} if there is an excess of AB haplotypes within the population ($D > 0$), and an increase if there is a deficit of AB

haplotypes within the population ($D < 0$). Our LD in the next generation is $D' = p'_{AB}$, so we can rewrite the above eqn. in terms of the D'

$$D' = (1 - r)D \tag{61}$$

so if the level of LD in generation 0 is D_0 the level t generations later (D_t) is

$$D_t = (1 - r)^t D_0 \tag{62}$$

so recombination is acting to decrease LD, and it does so geometrically at a rate given by $(1 - r)$. If $r \ll 1$ then we can approximate this by an exponential and say that

$$D_t \approx D_0 e^{-rt} \tag{63}$$

Q C) You find a hybrid population between the two mouse subspecies described in the question above, which appears to be comprised of equal proportions of ancestry from the two subspecies. You estimate LD between the two markers to be 0.0723. Assuming that this hybrid population is large and was formed by a single mixture event, can you estimate how long ago this population formed?

4 One locus models of selection

4.1 fitness

We will define the absolute fitness of a genotype to be the expected number of offspring of an individual of that genotype. Natural selection occurs when there are differences between our genotypes in their fitness. This difference could occur at any point during the life cycle.

4.2 Haploid selection model

The number of individuals carrying allele 1 and allele 2 in generation t are P_t and Q_t respectively. The current frequency of allele 1 is $p = P_t/(P_t + Q_t)$.

In the next generation the number of type 1 and 2 individual is $P_{t+1} = w_1 P_t$ and $Q_{t+1} = w_2 Q_t$ The mean fitness of our population is $w_1 p_t + w_2 q_t = \overline{w}$, i.e. the fitness of the two alleles weighted by their frequencies within the population.

The frequency of allele 1 in the next generation

$$p_{t+1} = \frac{w_1 P_t}{w_1 P_t + w_2 Q_t} = \frac{w_1 p_t}{\overline{w}} \tag{64}$$

The change in frequency from one generation to the next is

$$\Delta p = p_{t+1} - p_t = \frac{w_1 p_t}{\overline{w}} - p_t = \frac{pq(w_1 - w_2)}{\overline{w}} \tag{65}$$

As this fraction represents a ratio of fitnesses, we only need to specify our fitness up to an arbitrary constant. I.e. we can use relative fitnesses in this equation (65) as we are free to use $w_1/w_1 = 1$ and w_2/w_1 in place of w_1 and w_2 because as long as we use them consistently in the numerator and denominator of (65) the arbitrary constant $1/w_1$ will cancel out. Intuitively this makes sense, you can produce a huge number of children but you are out of luck as far as natural selection is concerned if others in the population are having more. What matters is your fitness relative to others in the population. By convention we do this by dividing through all our absolute fitnesses by that of the most fit individual, so that the fittest type in our population has a relative fitness of one.

Assuming that $w_1 > w_2$ our relative fitnesses are 1 and $w_2/w_1 < 1$ respectively, we will sometimes replace $w_2/w_1 = 1 - s$. Our s here is a selection coefficient the difference in relative fitnesses between our haploid alleles.

Assuming that the fitnesses of our two alleles are constant over time, the number of the 2 allelic types τ generations later is $P_{t+\tau} = (w_1)^\tau P_t$ and $Q_{t+\tau} = (w_2)^\tau Q_t$ and so

$$p_{t+\tau} = \frac{(w_1)^\tau P_t}{(w_1)^\tau P_t + (w_2)^\tau Q_t} = \frac{p_t}{p_t + q_t(1 - s)^\tau} \tag{66}$$

as $(w_2/w_1) = 1 - s$ then if $s \ll 1$

$$p_{t+\tau} \approx \frac{p_t}{p_t + q_t e^{-s\tau}} \tag{67}$$

This form is logistic growth, and follows from the fact that we are looking at the relative frequencies of two populations (allele 1 and 2) that are growing (or declining) exponentially.

Rearranging (66) we can work out the time for our frequency to change from a frequency p_0 to p' as follows

$$\frac{p'}{q'} = \frac{p_0}{q_0} \left(\frac{w_1}{w_2} \right)^t \tag{68}$$

therefore, using the fact that $w_1/w_2 = 1/(1 - s)$

$$-t \log(1 - s) = \log \left(\frac{p' \, q_0}{q' \, p_0} \right) \tag{69}$$

assuming that $s \ll 1$ we can replace the left hand side by ts.

One particular case of interest is the time it takes to go through introduction to near fixation in a population of size N (e.g. $p = 1/N$ to $p' = 1 - 1/N$) this takes time $t \approx \log(N)/s$ (assuming $s \ll 1$).

Haploid model with fluctuating selection   We can now consider the case where our fitnesses depend on time, and say that $w_{1,t}$ and $w_{2,t}$ are the fitnesses of the two types in generation t. The frequency of allele 1 in generation $t + 1$ is

$$p_{t+1} = \frac{w_{1,t} p_t}{\overline{w}_t} \tag{70}$$

The ratio of the frequency of allele 1 to allele 2 in generation $t + 1$ is

$$\frac{p_{t+1}}{q_{t+1}} = \frac{w_{1,t}}{w_{2,t}} \frac{p_t}{q_t} \tag{71}$$

Therefore if we think of our alleles starting in generation 0 at frequencies p_0 and q_0, then $t + 1$ generations later

$$\frac{p_{t+1}}{q_{t+1}} = \left(\prod_{i=0}^{t} \frac{w_{1,i}}{w_{2,i}} \right) \frac{p_t}{q_t} \tag{72}$$

So the question of which allele is increasing or decreasing in frequency comes down to whether $\left(\prod_{i=0}^{t} \frac{w_{1,i}}{w_{2,i}} \right)$ is > 1 or < 1. As it is a little hard to think about this ratio, we can instead take the tth root of this and instead consider

$$\sqrt[t]{\left(\prod_{i=0}^{t} \frac{w_{1,i}}{w_{2,i}} \right)} = \frac{\sqrt[t]{\prod_{i=0}^{t} w_{1,i}}}{\sqrt[t]{\prod_{i=0}^{t} w_{2,i}}} \tag{73}$$

24

$\sqrt[t]{\prod_{i=0}^{t} w_{1,i}}$ is the geometric mean fitness of allele 1 over our t generations. Therefore our allele 1 will only increase in frequency if it has a higher geometric mean fitness than allele 2 (at least in our simple deterministic model).

4.3 Diploid model

We'll move now to a diploid model of a single locus segregating 2 alleles. We'll assume that our difference in fitness between our three genotypes comes from differences in viability, i.e. differential survival of individuals of our three genotypes to reproduction. The fitnesses of three genotypes will be denoted by w_{11}, w_{12}, and w_{22}. On our individuals mate at random, so the number of our three genotypes at birth are:

$$Np_t^2, \quad N2p_tq_t, \quad Nq_t^2 \tag{74}$$

The mean fitness of the population is then

$$w_{11}p_t^2 + w_{12}2p_tq_t + w_{22}q_t^2 \tag{75}$$

We can now ask how many of each of our three genotypes survive to reproduce. Well an individual of genotype 11 has a probability of w_{11} of surviving to reproduce, and similarly for other genotypes. So that the number of our three genotypes who survive to reproduce is

$$Nw_{11}p_t^2, \quad Nw_{12}2p_tq_t, \quad Nw_{22}q_t^2 \tag{76}$$

it then follows that the total number of individuals who survive to reproduce is

$$N\left(w_{11}p_t^2 + w_{12}2p_tq_t + w_{22}q_t^2\right) \tag{77}$$

this is simply our mean fitness of the population multiplied by the population size (i.e. $N\overline{w}$).

The frequency of our 11 genotype individuals at reproduction is simply the number of 11 genotype individuals at reproduction ($Nw_{11}p_t^2$) divided by the total number of individuals who survive to reproduce ($N\overline{w}$), and likewise for our other two genotypes. Therefore, the frequency of individuals with the three different genotypes at reproduction is

$$\frac{Nw_{11}p_t^2}{N\overline{w}}, \quad \frac{Nw_{12}2p_tq_t}{N\overline{w}}, \quad \frac{Nw_{22}q_t^2}{N\overline{w}} \tag{78}$$

see Table 3.

	11	12	22
Num. at birth	Np_t^2	$N2p_tq_t$	Nq_t^2
Fitnesses	w_{11}	w_{12}	w_{22}
Num. at repro.	$Nw_{11}p_t^2$	$Nw_{12}2p_tq_t$	$Nw_{22}q_t^2$
freq. at repro.	$\frac{Nw_{11}p_t^2}{N\overline{w}}$	$\frac{Nw_{12}2p_tq_t}{N\overline{w}}$	$\frac{Nw_{22}q_t^2}{N\overline{w}}$

<div align="center">Table 3:</div>

As there is no difference in the fecundity of our three genotypes, the allele frequency in the offspring in the next generation is simply the allele frequency in the individuals. The frequency in the next generation is

$$p_{t+1} = \frac{w_{11}p_t^2 + w_{12}p_tq_t}{\overline{w}} \tag{79}$$

note that again the absolute value of our fitnesses is irrelevant to the frequency of the allele. Therefore, we can just as easily replace our absolute fitnesses with our relative fitnesses.

The change in frequency from generation t to $t+1$ is

$$\Delta p = p_{t+1} - p_t = \frac{w_{11}p_t + w_{12}p_tq_t}{\overline{w}} - p_t. \tag{80}$$

To simplify this equation we will first define two variables $\overline{w}_1$ and $\overline{w}_2$ as

$$\overline{w}_1 = w_{11}p_t + w_{12}q_t, \tag{81}$$
$$\overline{w}_2 = w_{12}p_t + w_{22}q_t. \tag{82}$$

Our variables $\overline{w}_1$ and $\overline{w}_2$ are called the marginal fitnesses of allele 1 and allele 2 respectively. They are called this as $\overline{w}_1$ is the average fitness of an allele 1, i.e. the fitness of the 1 allele in a homozygote weighted by the probability it is in a homozygote (p_t) plus the fitness of the 1 allele in a heterozygote weighted by the probability it is in a heterozygote (q_t). We can then rewrite (80) usng $\overline{w}_1$ and $\overline{w}_2$ as

$$\Delta p = \frac{p_tq_t(\overline{w}_1 - \overline{w}_2)}{\overline{w}}. \tag{83}$$

The sign of Δp, i.e. whether allele 1 increases of decreases in frequency, depends only on the sign of $(\overline{w}_1 - \overline{w}_2)$. The frequency of allele 1 will keep increasing over the generations so long as its marginal fitness is higher than that of allele 2, i.e. $\overline{w}_1 > \overline{w}_2$, while if $\overline{w}_1 < \overline{w}_2$ then the frequency of allele 1 will decrease. (We will return to the special case where $\overline{w}_1 = \overline{w}_2$ shortly).

We can also rewrite (80) as

$$\Delta p_t = \frac{p_tq_t}{\overline{w}} \frac{d\overline{w}}{dp}. \tag{84}$$

the demonstration of this we leave to the reader. This form shows that Δp in increase if $\frac{d\overline{w}}{dp} > 1$, i.e. increasing the frequency of 1 increases the mean fitness, while the frequency of the allele with decrease if this increases the mean fitness of the population ($\frac{d\overline{w}}{dp} > 1$). Thus although selection acts on individuals, under this simple model selection is acting to increase the mean fitness of the population, and it does so at a rate given by the variance in allele frequencies within the population (pq).

Question) Show that (84) and (80) are equivalent.

4.3.1 Diploid directional selection

Our diploid model is going to reveal a number of insights. We'll start with a simple model of directional selection, i.e. one our our alleles is always has higher marginal fitness than the other. Let's assign allele 1 to be the fitter allele, so that $w_{11} \geq w_{12} \geq w_{22}$. As we are interested in changes in allele frequencies we are only interested in relative fitnesses. To parameterize our reduction in relative fitness in terms of a selection coefficient, similar to the one we met in the haploid selection section, as follows

	11	12	22
abs. fitness	w_{11}	$\leq w_{12} \leq$	w_{22}
rel. fitness	w_{11}/w_{11}	w_{12}/w_{11}	w_{22}/w_{11}
rel. fitness	1	$1 - sh$	$1 - s$

here our selection coefficient s is the difference in relative fitness between our two homozygotes, while we will call h our dominance coefficient. Our dominance coefficient allows us to move from the situation where allele 1 has a fully dominant effect on fitness (and 2 is totally recessive) when $h = 0$, such that the heterozygote exactly resembles the 1 homozygotes, to the converse case where the allele 1 is fully recessive ($h = 1$, and allele 2 is dominant).

We can then rewrite (83) as

$$\Delta p = \frac{p_t q_t (p_t h s + q s (1 - h))}{\overline{w}}. \tag{85}$$

where

$$\overline{w} = 1 - 2 p_t q_t s h - q_t^2 s \tag{86}$$

One special case is when $s_{12} = s_{22}/2$

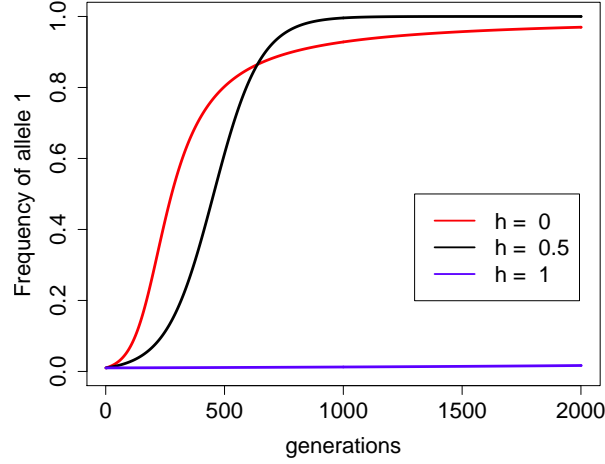$$\Delta p = \frac{p_t q_t s/2}{\overline{w}}. \tag{87}$$

27

Figure 3: The trajectory of Allele 1 through the population starting from $p = 0.01$ for a selection coefficient $s = 0.01$ and three different dominance coefficients.

for values of $s \ll 1$ then the denominator is close to 1 and this is of exactly the same form as our haploid model and so if s is constant our trajectory follows a logistic growth curve of the form (67). Thus we can show that it takes

$$\approx 2log(2N)/s \tag{88}$$

generations for our selected allele to transit from its entry into the population ($p_0 = 1/(2N)$) to close to fixation ($p' = 1 - 1/(2N)$).

4.3.2 heterozygote advantage

What about the case where our heterozygotes are fitter than either of the homozygotes. We parameterize the relative fitnesses as follows

	11	12	22
abs. fitness	w_{11}	$< w_{12} >$	w_{22}
rel. fitness	w_{11}/w_{12}	w_{12}/w_{12}	w_{22}/w_{12}
rel. fitness	$1 - s_1$	1	$1 - s_2$

so s_1 and s_2 are the differences between the relative fitnesses of the two homozygotes and the heterozygote. Note that to obtain relative fitnesses we've divided through our fitnesses by the heterozygote fitness. We could use the same parameterization as in our directional selection model, but this slight reparameterization makes the math prettier.

In this case, when our allele 1 is rare it is often found in a heterozygous state and so it increases in frequency. However, when allele 1 is common it is often found in the homozy-

28

gote state, while the allele 2 is often found in the heterozygote state, and so it is now 2 that increases in frequency at the expense of allele 1. Thus, at least in our deterministic model, neither allele can reach fixation and our allele will be maintained as a balanced polymorphism in the population at an equilibrium frequency.

We can solve for this equilibrium frequency by setting $\Delta p = 0$ in (83), i.e. $p_t q_t (\overline{w}_1 - \overline{w}_2) = 0$ which permits three stable equilibrium, two uninteresting ($p = 0$ or $q = 0$) and one polymorphic where $\overline{w}_1 - \overline{w}_2$ at equilibrium frequency p_e. In this case the marginal fitnesses of our two alleles are exactly equal ($\overline{w}_1 - \overline{w}_2$), substituting in our selection coefficients this implies

$$p_e = \frac{s_2}{s_1 + s_2} \tag{89}$$

This is also that the mean fitness of the population is maximised.

Under-dominance. Another case that is of potential interest is the case of fitness under-dominance where the heterozygote is less fit than either of the homozygotes.

	11	12	22
abs. fitness	w_{11}	$> w_{12} <$	w_{22}
rel. fitness	$1 + s_1$	1	$1 + s_2$

this case also permits three equilibria $p = 0$, $p = 1$, and a polymorphic equilibrium $p = p_U$ however now only the first two equilibria are stable, while the polymorphic equilibrium is now unstable. If $p < p_U$ then Δp is negative and the allele 1 proceeds to loss, while if $p > p_U$ then the allele 1 will proceed to fixation.

While such alleles might not spread within populations (if $p_U \gg 0$ and selection is reasonably strong), they are of interest in the study of speciation and hybrid zones. That's because our allele 1 and 2 may have arisen in a stepwise fashion, i.e. not by a single mutation, in separate subpopulations and our heterozygote disadvantage will now play a potential role in species maintenance.

Diploid Fluctuating fitness We would like to think about the case where the diploid absolute fitnesses are time-dependent with our three genotypes having fitnesses $w_{11,t}$, $w_{12,t}$, and $w_{22,t}$ in generation t. However, this case is much less tractable than the haploid case, as segregation makes it tricky to keep track of the genotype frequencies. We can make some progress and gain some intuition by thinking about how the frequency of allele 1 changes when it is rare. (This argument is originally due to Haldane and J.)

When allele 1 is rare, i.e. $p \ll 1$ our frequency in the next generation (79) can be approximated as

$$p_{t+1} \approx \frac{w_{12} p_t}{\overline{w}} \tag{90}$$

by ignoring the p_t^2 term and assuming $q_t \approx 1$ in the numerator. Following a similar argument to approximate q_{t+1}, we can write

$$\frac{p_{t+1}}{q_{t+1}} = \frac{w_{12,t}}{w_{22,t}} \frac{p_t}{q_t} \tag{91}$$

Then starting from out from p_0 and q_0 in generation 0, $t+1$ generations later

$$\frac{p_{t+1}}{q_{t+1}} = \left(\prod_{i=1}^{t} \frac{w_{12,i}}{w_{22,i}} \right) \frac{p_0}{q_0}. \tag{92}$$

From this we can see, following our haploid argument from above, that the frequency of our allele 1 will increase when rare only if

$$\frac{\sqrt[t]{\prod_{i=0}^{t} w_{12,i}}}{\sqrt[t]{\prod_{i=0}^{t} w_{22,i}}} > 1 \tag{93}$$

i.e. if the 12 heterozygote has higher geometric mean fitness than the 22 homozygote.

The question now is, will allele 1 approach fixation in the population, or are there cases where our allele can become a balanced polymorphism? To investigate that we can simply repeat our analysis for $q \ll 1$, and see that in that case

$$\frac{p_{t+1}}{q_{t+1}} = \left(\prod_{i=1}^{t} \frac{w_{11,i}}{w_{12,i}} \right) \frac{p_0}{q_0}. \tag{94}$$

i.e. now for our allele 1 to carry on increasing in frequency and to approach fixation in our population our 11 genotype has to be out-completing our 12 heterozygotes. For our allele 1 to approach fixation we need the geometric mean of $w_{11,i}$ to be greater than the geometric mean fitness of heterozygotes ($w_{12,i}$) While if heterozygotes have higher geometric mean fitness than our 1 homozygotes than the 2 allele will increase in frequency when it is rare. Therefore, a balanced polymorphism can result when the heterozygote has higher geometric fitness than either of the homozygotes.

Intriguingly we can have a balanced polymorphism even if the heterozygote is never the fittest genotype in any generation. To see this consider the simple example, where there are two environments alternate generation to generations

	11	12	22
Environ. A abs. fitness	$w_{11,A}$	$> w_{12,A} >$	$w_{22,A}$
Environ. B abs. fitness	$w_{11,B}$	$< w_{12,B} <$	$w_{22,B}$
Geometric mean fitness	$w_{11,B}$	$< w_{12,B} >$	w_{22}

so that the polymorphism will remain balanced in the population, despite the fact that the heterozygote is never the fitest genotype.

4.4 Mutation Selection Balance

Mutation is constantly introducing new alleles into the population. Therefore, variation can be maintained within our population even if it is deleterious through a balance between mutation introducing alleles and selection acting to purge these alleles. To see this we'll return to our directional selection model, where the allele 1 is advantageous i.e.

$$
\begin{array}{cccc}
 & 11 & 12 & 22 \\
\text{abs. fitness} & w_{11} & \geq w_{12} \geq & w_{22} \\
\text{rel. fitness} & 1 & 1 - sh & 1 - s
\end{array}
$$

and for the moment we'll consider the case where the allele 2 is not completely recessive $h > 0$. We'll be interested in the case where our allele 2 is rare within the population i.e. $q \ll 1$, which means that the change in frequency of allele 2, due to selection ($\Delta_S q$) can be written as

$$\Delta_S q = \frac{pq(\overline{w}_2 - \overline{w}_1)}{\overline{w}} \approx -hsq \tag{95}$$

which can be found by assuming that $q^2 \approx 0$, $p \approx 1$, and that $\overline{w} \approx w_1$. So selection is acting to reduce the frequency of our allele 2 and does so geometrically across the generations.

We'll now consider the change in frequency induced by mutation. Lets assume that our mutation rate from allele 1 to allele 2 is μ per generation (mutation also could occur from allele 2 to 1 however this is a small effect as long as the reverse mutation rate isn't too large). The frequency of q in the next generation is

$$q' = \mu(1 - q) \tag{96}$$

assuming that $\mu \ll 1$ and that $q \ll 1$ then the change in the frequency of allele 2 due to mutation ($\Delta_M q$) is

$$\Delta_M q = \mu \tag{97}$$

i.e. mutation is, when allele 2 is rare, is acting to linearly increase the frequency of the deleterious allele.

We'll obtain a balance between mutation acting to increase the frequency of our allele and selection acting to decrease the frequency of our allele when these two forces are at equilibrium i.e.

$$\Delta_M q + \Delta_S q = 0 \tag{98}$$

this equilibrium occurs when

$$q_e = \frac{\mu}{hs} \tag{99}$$

i.e. our allele frequency is balanced at our mutation rate divided by the reduction in relative fitness in the heterozygote (hs).

Note that in this calculation the fitness of the 22 homozygote hasn't entered into this, as our allele is rare and so rarely in a homozygous state. Therefore, if our allele has any deleterious effect in a heterozygous state it is this effect that determines the frequency it is maintained at within the population. Note that in writing our total allele frequency change as $\Delta_M q + \Delta_S q$ we have implicitly assumed that we can ignore terms of order $\mu \times s$ as they are small (this allows us to separate the change in allele frequencies from mutation and selection).

So what effect do such mutations have on the population? Consider the effect a single site segregating at $q_E = \mu/(hs))$ has on mean relative fitness

$$\overline{w} = 1 - 2p_E q_E hs - q_E^2 s \approx 1 - 2u \tag{100}$$

somewhat remarkably the drop in mean fitness due to a site segregating at mutation selection balance is (unless the site is totally recessive) independent of the selection coefficient against the heterozygote and depends only the mutation rate.

While this reduction is very small at an individual site (e.g. the mutation rate of a gene is likely $< 1^{-5}$) there are many loci segregating at mutation selection balance. This can be a big source of genetic load and a major cause of variation in fitness related traits among individuals.

As an aside, if an allele was truly recessive (although few likely are) $h = 0$ and so (99) is not valid, however, we can proceed through a similar line of reasoning, to that outlined above, to show that for truly recessive alleles $q_e = \sqrt{\mu/s}$.

4.4.1 Inbreeding depression

All else being equal, mutations that have a smaller effect in the heterozygote can segregate at higher frequency under mutation selection balance. As a consequence of this, alleles that have strongly deleterious effects in the homozygous state can segregate at low frequencies in the population, as long as they do not have a strong effect in heterozygotes. Thus outbred populations may have many alleles with recessive deleterious effects segregating within them.

On consequence of this is that inbred individuals from usually outbred populations may have dramatically lower fitnesses than outbred individuals as a consequence of being homozygous at many loci for alleles with recessive deleterious effects. Indeed this seems to be the case a common observation (dating back to systematic surveys by Darwin) in that in typically outbred populations the mean fitness of individuals decreases with the inbreeding coefficient, i.e. inbreeding depression is a common observation.

Purging the inbreeding load That said, populations who regularly inbreed over sustained time periods are expected to partially purge this load of deleterious alleles. This is because these populations have exposed many of these alleles in a homozygous state and so selection can more readily remove these alleles from the population.

4.5 Migration-Selection Balance

Another reason for the persistence of deleterious alleles in a population is that there is a constant influx of maladaptive alleles from other populations where these alleles are locally adapted. This seems unlikely to be as broad an explanation for the persistence of deleterious alleles genome-wide as mutation-selection balance. However, a brief discussion of such alleles is worthwhile as it helps to inform our ideas about local adaptation.

As a first pass at this lets consider a haploid two allele model with two different populations, where the relative fitnesses of our alleles are as follows

allele	1	2
population 1	1	1-s
population 2	1-s	1

As a simple model of migration lets suppose within a population a fraction of m individuals are migrants from the other population, and $1 - m$ individuals are from the same deme.

To quickly sketch a solution to this well set up a situation analogous to our mutation-selection balance model. to do this lets assume that selection is strong compared to migration ($s \gg m$) then allele 1 will be almost fixed in population 1 and allele 2 will be almost fixed in population 2. If that is the case, migration changes the frequency of allele 2 in population 1 (q_1) by

$$\Delta_{Mig.}q_1 \approx m \tag{101}$$

while as noted above $\Delta_S q_1 = -sq_1$, so that migration and selection are at an equilibrium when $\Delta_S q_1 + \Delta_{Mig.}q_1$, i.e. an equilibrium frequency of allele 2 in population 1 of

$$q_{e,1} = \frac{m}{s} \tag{102}$$

so that migration is playing to role of mutation and so migration-selection balance (at least under strong selection) is analogous to mutation selection balance.

4.5.1 Some theory of the spatial distribution of allele frequencies under deterministic models of selection

Imagine a continuous haploid population spread out along a line. individual dispersals a random distance Δx from its birthplace to the location where it reproduces, where Δx is drawn from the probability density $g(\ )$. To make life simple we will assume that $g(\Delta x)$ is normally distributed with mean zero and standard deviation σ, i.e. migration is unbiased an individuals migrate an average distance of σ.

Our frequency of allele 2 at time t in the population at spatial location x is $q(x, t)$. Assuming that only dispersal occurs, how does our allele frequency change in the next

generation. Our allele frequency in the next generation at location x reflects the migration from different locations in the proceeding generation. Our population at location x receives a contribution $g(\Delta x)q(x + \Delta x, t)$ of allele 2 from the population at location $x + \Delta x$, such that the frequency of our allele at x in the next generation is

$$q(x, t+1) = \int_{-\infty}^{\infty} g(\Delta x)q(x + \Delta x, t)d\Delta x. \tag{103}$$

To obtain $q(x + \Delta x, t)$ lets take a taylor series expansion of $q(x, t)$

$$q(x + \Delta x, t) = q(x, t) + \Delta x \frac{dq(x, t)}{dx} + \tfrac{1}{2}(\Delta x)^2 \frac{d^2 q(x, t)}{dx^2} + \cdots \tag{104}$$

then

$$q(x, t+1) = q(x, t) + \left(\int_{-\infty}^{\infty} \Delta x g(\Delta x) d\Delta x \right) \frac{dq(x, t)}{dx} + \tfrac{1}{2} \left(\int_{-\infty}^{\infty} (\Delta x)^2 g(\Delta x) d\Delta x \right) \frac{d^2 q(x, t)}{dx^2} + \cdots \tag{105}$$

$g(\ )$ has a mean of zero so $\int_{-\infty}^{\infty} \Delta x g(\Delta x) d\Delta x = 0$ and has variance σ^2 so $\int_{-\infty}^{\infty} (\Delta x)^2 g(\Delta x) d\Delta x = \sigma^2$ and all higher terms are zero (as all high moments of the normal are zero). Looking at the change in frequency $\Delta q(x, t) = q(x, t+1) - q(x, t+1)$ then

$$\Delta q(x, t) = \frac{\sigma^2}{2} \frac{d^2 q(x, t)}{dx^2} \tag{106}$$

this is a diffusion equation, so that migration is acting to smooth out allele frequency differences with a diffusion constant of $\frac{\sigma^2}{2}$. This is exactly analogous to the equation describing how a gas diffuses out to equal density, as both particles in a gas and our individuals of type 2 are performing brownian motion (blurring our eyes and seeing time as continuous).

We will now introduce fitness differences into our model and set the relative fitnesses of allele 1 and 2 at location x to be 1 and $1 + s\gamma(x)$. To make progress in this model we'll have to assume that selection isn't too strong i.e. $s\gamma(x) \ll 1$ for all x. The the change in frequency of allele 2 obtained within a generation due to selection is

$$q'(x, t) - q(x, t) \approx s\gamma(x)q(x, t)\big(1 - q(x, t)\big) \tag{107}$$

i.e. logistic growth of our favoured allele at location x. Putting our selection and migration terms together we find

$$q(x, t+1) - q(x, t) = s\gamma(x)q(x, t)\big(1 - q(x, t)\big) + \frac{\sigma^2}{2} \frac{d^2 q(x, t)}{dx^2} \tag{108}$$

in derving this we have essentially assumed that migration acted upon our original frequencies before selection and in doing so have ignored terms of the order of σs.

To make progress lets consider a simple model of location adaptation where the environment abruptly changes. Specifically we assume that $\gamma(x) = -1$ for $x < 0$ and $\gamma(x) = 1$
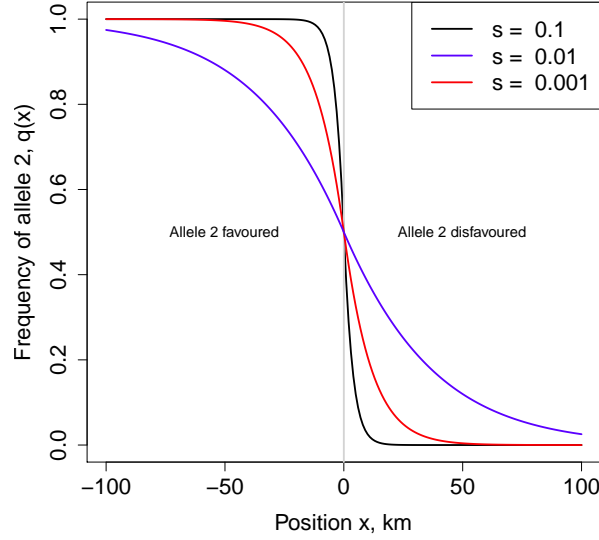
Figure 4: An equilibrium cline in allele frequency. Our individuals dispersal an average distance of $\sigma = 1$km per generation, and our allele 2 has a relative fitness of $1 + s$ and $1 - s$ on either side of the environmental change at $x = 0$.

for $x \geq 0$, i.e. our allele 2 has a selective advantage at locations to the left of zero, while this allele is at a disadvantage to the right of zero. In this case we can get an equilibrium distribution of our two alleles were to the left of *zero* our allele 2 is at higher frequency, while to the right of zero allele 1 predominates. As we cross from the left to the right side of our range the frequency of our allele 2 decreases in a smooth cline.

Our equilibrium spatial distribution of allele frequencies can be found by setting the LHS of eqn. (108) to zero to arrive at

$$s\gamma(x)q(x)\left(1 - q(x)\right) = \frac{\sigma^2}{2}\frac{d^2q(x)}{dx^2} \tag{109}$$

We then could solve this differential equation with appropriate boundary conditions ($q(-\infty) = 1$ and $q(\infty) = 0$) to arrive at the appropriate functional form to our cline. While we won't go into the solution of this equation here, we can note that by dividing our distance x by $\ell = \sigma/\sqrt{s}$ we can remove the effect of our parameters from the above equation. This compound parameter ℓ is the characteristic length of our cline, and it is this parameter which determines over what geographic scale we change from allele 2 predominating to allele 1 predominating as we move across our environmental shift.

The width of our cline, i.e. over what distance do we make this shift from allele 2 predominating to allele 1, can be defined in a number of different ways. One simple way to define the

cline width, which is easy to define but perhaps hard to measure accurately, is the slope (i.e. the tangent) of $q(x)$ at $x = 0$. Under this definition the cline width is approximately $0.6\sigma/\sqrt{s}$.

5 Stochasticity and Genetic Drift in allele frequencies

5.1 Stochastic loss of strongly selected alleles

Even strongly selected alleles can be lost from the population when they are sufficiently rare. This is because the number of offspring left by individuals to the next generation is fundamentally stochastic. A selection coefficient of s=1% is a strong selection coefficient, which can drive an allele through the population in a few hundred generations once the allele is established. However, if individuals have on average a small number of offspring per generation the first individual to carry our allele who has on average 1% more children could easily have zero offspring, leading to the loss of our allele before it ever get a chance to spread.

To take a first stab at this problem lets think of a very large haploid population, and in order for this population to stay constant in size we'll assume that individuals without the selected mutation have on average one offspring per generation. While individuals with our selected allele have on average $1 + s$ offspring per generation. We'll assume that the distribution of offspring number of an individual is Poisson distributed with this mean, i.e. the probability that an individual with the selected allele has i children is

$$P_i = \frac{(1 + s)^i e^{-(1+s)}}{i!} \tag{110}$$

Consider starting from a single individual with the selected allele, and ask about the probability of eventual loss of our selected allele starting from this single copy (p_L). To derive this we'll make use of a simple argument (derived from branching processes). Our selected allele will be eventually lost from the population if every individual with the allele fails to leave descendents.

1. In our first generation with probability P_0 our individual leaves no copies of itself to the next generation, in which case our allele is lost.

2. Alternatively it could leave one copy of itself to the next generation (with probability P_1), in which case with probability p_L this copy eventually goes extinct.

3. It could leave two copies of itself to the next generation (with probability P_2), in which case with probability p_L^2 both of these copies eventually goes extinct.

4. More generally it could leave could leave k copies ($k > 0$) of itself to the next generation (with probability P_k), in which case with probability p_L^k all of these copies eventually go extinct.

summing over this probabilities we see that

$$
\begin{aligned}
p_L &= \sum_{k=0}^{\infty} P_k p_L^k \\
&= \sum_{k=0}^{\infty} \frac{(1+s)^k e^{-(1+s)}}{k!} p_L^k \\
&= e^{-(1+s)} \left(\sum_{k=0}^{\infty} \frac{(p_L(1+s))^k}{k!} \right)
\end{aligned} \tag{111}
$$

well the term in the brackets is itself an exponential expansion, so we can rewrite this as

$$p_L = e^{(1+s)(p_L-1)} \tag{112}$$

solving this would give us our probability of loss for any selection coefficient. Lets rewrite this in terms of the the probability of escaping loss $p_F = 1 - p_L$. We can rewrite eqn (112) as

$$1 - p_F = e^{-p_F(1+s)} \tag{113}$$

to gain an approximation to this lets consider a small selection coefficient $s \ll 1$ such that $p_F \ll 1$ and then expanded out the exponential on the right hand side (ignoring terms of higher order than s^2 and p_F^2) then

$$1 - p_F \approx 1 - p_F(1+s) + p_F^2(1+s)^2/2 \tag{114}$$

solving this we find that

$$p_F = 2s. \tag{115}$$

Thus even an allele with a 1% selection coefficient has a 98% probability of being lost when it is first introduced into the population by mutation.

We can also adapt this result to a diploid setting. Assuming that heterozygotes for the 1 allele have $1 + (1-h)s$ children, the probability of allele 1 is not lost, starting from a single copy in the population, is

$$p_F = 2(1-h)s \tag{116}$$

for $h > 0$.

5.2 The interaction between genetic drift and weak selection.

For strongly selected alleles, once the allele has escaped initial loss at low frequencies, their path will be determined deterministically by their selection coefficients. However, if selection is weak the stochasticity of reproduction can play a role in the trajectory an allele takes even when it is common in the population.

To see this lets think of our simple Wright-Fisher model (see R exercise). Each generation we allow a deterministic change in our allele frequency, and then binomially sample two alleles for each of our offspring to construct our next generation.

So the expected change in our allele frequency within a generation is given just by our deterministic formula. To make things easy on our self lets assume an additive model, i.e. $h = 1/2$, and that $s \ll 1$ so that $\overline{w} \approx 1$. This gives us

$$\mathbb{E}(\Delta p) = \frac{s}{2}p(1-p) \tag{117}$$

our variance in our allele frequency change is given by

$$Var(p' - p) = Var(p') = \frac{p'(1 - p')}{2N} \tag{118}$$

this variance in our allele frequency follows from the fact that we are binomially sampling $2N$ new alleles in the next generation from a frequency p'. Denoting our count of allele 1 by i our

$$Var(p' - p) = Var(\frac{i}{2N} - p) = Var(\frac{i}{2N}) = \frac{Var(i)}{(2N)^2} \tag{119}$$

and from binomial sampling $Var(i) = 2Np'(1-p')$ and so we arrive at our answer. Assuming that $s \ll 1$, $p' \approx p$, then in practice we can use

$$Var(\Delta p) = Var(p' - p) \approx \frac{p(1 - p)}{2N}. \tag{120}$$

To get our first look at the relative effects of selection vs drift we can simply look at when our change in allele frequency caused selection within a generate is reasonably faithfully passed across the generations. In particular if our expected change in frequency is much great than the variance around this change, genetic drift will play little role in the fate of our selected allele (once the allele is not too rare within the population). When does selected dominant genetic drift? This will happen if $\mathbb{E}(\Delta p) \gg Var(\Delta p)$ when $Ns \gg 1$. Conversely any hope of our selected allele following its deterministic path will be quickly undone if our change in allele frequencies due to selection is much less than the variance induced by drift. So if $Ns \ll 1$ then drift will dominate the fate of our allele.

To make further progress on understanding the fate of alleles with selection coefficients of the order $1/N$ requires more careful modeling. However, we can obtain the probability that under our diploid model, with an additive selection coefficient s, the probability of allele 1 fixing within the population starting from a frequency p is given by

$$\pi(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}} \tag{121}$$

the proof of this is sketched out below (see Section 5.2.2). A new allele will arrive in the population at frequency $p = 1/(2N)$, then its probability of reaching fixation is

$$\pi\left(\frac{1}{2N}\right) = \frac{1 - e^{-s}}{1 - e^{-2Ns}} \tag{122}$$

if $s \ll 1$ but $Ns \gg 1$ then $\pi(\frac{1}{2N}) \approx s$, which nicely gives us back our result that we obtained above (eqn. (116)).

In the case where Ns close to 1 then

$$\pi(\frac{1}{2N}) \approx \frac{s}{1 - e^{-2Ns}} \tag{123}$$

this is greater than s, increasingly so for smaller N. Why is this? Well in small populations selected alleles spend a somewhat shorter time segregating (especially at low frequencies), and so are slightly less susceptible to genetic drift.

5.2.1 The fixation of slightly deleterious alleles.

We can also use eqn. (123) to understand how likely it is that deleterious alleles accidently reach fixation by genetic drift, assuming a diploid model with additive selection (with a selection coefficient of $-s$ against our allele 2). If $Ns \gg 1$ then our deleterious allele (allele 2) can not possibly reach fixation. However, if Ns is not large then

$$\pi(\frac{1}{2N}) = \frac{s}{e^{2Ns} - 1} \tag{124}$$

for our deleterious allele. So deleterious alleles can fix within populations (albeit at a low rate) if Ns is not too large.

5.2.2 A Sketch Proof of the probability of fixation of weakly selected alleles

We'll let $P(\Delta p)$ be the probability that our allele frequency shifts by Δp in the next generation. Using this we can write our probability $\pi(p)$ in terms of the probability of achieving fixation averaged over the frequency in the next generation

$$\pi(p) = \int \pi(p + \Delta p)P(\Delta p)d(\Delta p) \tag{125}$$

This is very similar to the technique that we used deriving our probability of escaping loss in a very large population above.

So we need an expression for $\pi(p+\Delta p)$. To obtain this we'll do a Taylor series expansion of $\pi(p)$ assuming that Δp is small

$$\pi(p + \Delta p) \approx \pi(p) + \Delta p \frac{d\pi(p)}{dp} + (\Delta p)^2 \frac{d^2\pi(p)}{dp^2}(p) \tag{126}$$

ignoring higher order terms.

Taking the expectation over Δp on both sides, as in eqn. 125, we obtain

$$\pi(p) = \pi(p) + \mathbb{E}(\Delta p)\frac{d\pi(p)}{dp} + \mathbb{E}((\Delta p)^2)\frac{d^2\pi(p)}{dp^2} \tag{127}$$

Well $\mathbb{E}(\Delta p) = \frac{s}{2}p(1-p)$ and $Var(\Delta p) = \mathbb{E}(\Delta p)^2) - \mathbb{E}^2(\Delta p)$, so if $s \ll 1$ then $\mathbb{E}^2(\Delta p) \approx 0$, and $\mathbb{E}(\Delta p)^2 = \frac{p(1-p)}{2N}$. This leaves us with

$$0 = \frac{s}{2}p(1-p)\frac{d\pi(p)}{dp} + \frac{p(1-p)}{2N}\frac{d^2\pi(p)}{dp^2} \tag{128}$$

and we can specify the boundary conditions to be $\pi(1) = 1$ and $\pi(0) = 0$. Solving this differential equation is somewhat involved process but in doing so we find that

$$\pi(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}} \tag{129}$$

6 Genetic drift and Neutral alleles

A neutral polymorphism occurs when the segregating alleles at a polymorphic site have no discernable effect on the fitness of the organism (i.e. if they have any effect on fitness at all drift overpowers it $Ns \ll 1$).

6.1 The fixation of neutral alleles

It is very unlikely that a rare neutral allele accidentally drifts up to fixation, it is much more likely that such an allele is eventually is lost from the population. However, there is a large and constant influx of rare alleles into the population due to mutation, so even if it is very unlikely that an individual allele fixes within the population, some neutral alleles will fix.

Probability of the eventual fixation of a neutral allele. An allele which reaches fixation within a population, is an ancestor to the entire population. In a particular generation there can be only single allele that all other alleles at the locus in later generation can claim as a ancestor. As at a neutral locus all of our alleles are exchangeable, as they have no effect on the number of descendents an individual leaves, so any allele is equally likely to be the ancestor of the entire population. In a diploid population size of size N, there are $2N$ alleles all of which are equally likely to be the ancestor of the entire population at some later time point. So if our allele is present in a single copy, the chance that is the ancestor to the entire population in some future generation is $1/(2N)$, i.e. the chance our neutral allele is eventually fixed is $1/(2N)$.

More generally if our neutral allele is present in i copies in the population, of $2N$ alleles, the probability that this allele is fixed is $i/(2N)$. I.e. the probability that a neutral allele is eventually fixed is simply given by its frequency (p) in the population. We can also derive this result by letting $Ns \to 0$ in eqn. (121).

Rate of substitution of neutral alleles. A substitution between populations that do not exchange gene flow is simply a fixation event within one population. The rate of substitution is therefore the rate at which new alleles fix in the population, so that the long-term substitution rate is the rate at which mutations arise that will eventually become fixed within our population.

Assume that there are two classes of mutational changes that can occur with a region, highly deleterious mutations and neutral mutations. A fraction C of all mutational changes are highly deleterious, and can not possibly contribute to substitution nor polymorphism (i.e. $Ns \gg 1$). While a fraction $1 - C$ are neutral. If our mutation rate is μ per transmitted allele per generation, then a total of $2N\mu(1 - C)$ neutral mutations enter our population each generation.

Each of these neutral mutations has a $1/(2N)$ probability chance of eventually becoming fixed in the population. Therefore, the rate at which neutral mutations arise that eventually become fixed within our population is

$$2N\mu(1 - C)\frac{1}{2N} = \mu(1 - C) \tag{130}$$

thus the rate of substitution under a model where newly arising alleles are either highly deleterious or neutral, is simply given by the mutation rate towards neutral alleles, i.e. $\mu(1-C)$.

Consider a pair of species have diverged for T generations, i.e. orthologous sequences shared between the species last shared a common ancestor T generations ago. If they have maintained a constant μ over that time, will have accumulated an average of

$$2\mu(1 - C)T \tag{131}$$

neutral substitutions. This assumes that T is a lot longer than the time it takes to fix a neutral allele, such that the total number of alleles introduced into the population that will eventually fix is the total number of substitutions. We'll see below that a neutral allele takes on average $4N$ generations to fix from its introduction into the population.

This is a really pretty result as the population size has completely canceled out of the neutral substitution rate. However, there is another way to see this in a more straightward way. If I look at a sequence in me compared to say a particular chimp, I'm looking at the mutations that have occurred in both of our germlines since they parted ways T generations ago. Since neutral alleles do not alter the probability of their transmission to the next generation, we are simply looking at the mutations that have occurred in $2T$ generations worth of transmissions. Thus the average number of neutral mutational differences separating our pair of species is simply $2\mu(1 - C)T$.

6.2 Loss of heterozygosity due to to drift.

Genetic drift will, in the absence of new mutations, slowly purge our population of genetic diversity as alleles slowly drift to high or low frequencies and are lost or fixed over time.

Imagine a population of a constant size N diploid individuals, and that we are examining a locus segregating for two alleles that are neutral with respect to each other. This population is randomly mating with respect to the alleles at this locus.

In generation t our current level of heterozygosity is H_t, i.e. the probability that two randomly sampled alleles in generation t are non-identical is H_t. Assuming that the mutation rate is zero (or vanishing small), what is our level of heterozygosity in generation $t+1$?

In the next generation $(t+1)$ we are looking at the alleles in the offspring of generation t. If we randomly sample two alleles in generation $t+1$ which had different parental alleles in generation t then it is just like drawing two random alleles from generation t. So the probability that these two alleles in generation $t+1$, that have different parental alleles in generation t, are non-identical is H_t.

Conversely, if our pair of alleles have the same parental allele in the proceeding generation (i.e. the alleles are identical by descent one generation back) then these two alleles must be identical (as we are not allowing for any mutation).

In a diploid population of size N individuals there are $2N$ alleles. The probability that our two alleles have the same parental allele in the proceeding generation is $1/(2N)$, the probability that they have different parental alleles is is $1 - 1/(2N)$. So by the above argument the expected heterozygosity in generation $t+1$ is

$$H_{t+1} = \frac{1}{2N} \times 0 + \left(1 - \frac{1}{2N}\right) H_t \tag{132}$$

By this argument if the heterzygosity in generation 0 is H_0 our expected heterozygosity in generation t is

$$H_{t+1} = \left(1 - \frac{1}{2N}\right)^t H_0 \tag{133}$$

i.e. the expected heterozygosity with our population is decaying geometrically with each passing generation. If we assume that $1/(2N) \ll 1$ then we can approximate this geometric decay by an exponential decay, such that

$$H_{t+1} = H_0 \exp\left(-\frac{t}{2N}\right) \tag{134}$$

i.e. heterozygosity decays exponentially at a rate $1/(2N)$.

6.3 Levels of diversity maintained by a balance between mutation and drift.

heterozygosity Looking backwards in time from one generation to the next, we are going to say that two alleles which have the same parental allele (i.e. find their common ancestor)

in the preceding generation have *coalesced*, and refer to this event as a *coalescent event*.

The probability that our pair of randomly sampled alleles have coalesced in the preceding generation is $1/(2N)$, the probability that our pair of alleles fail to coalesce is $1 - 1/(2N)$.

The probability that a mutation changes the identity of the transmitted allele is μ per generation. So the probability of no mutation occurring is $(1 - \mu)$. We'll assume that when a mutation occurs it creates some new allelic type which is not present in the population. This assumption (commonly called the infinitely-many-alleles model) makes the math slightly cleaner, and also is not too bad an assumption biologically.

This model lets us calculate when our two alleles last shared a common ancestor and whether these alleles are identical as a result of failing to mutate since this shared ancestor. For example we can work out the probability that our two randomly sampled alleles coalesced 2 generations in the past (i.e. they fail to coalesce in generation 1 and then coalescent in generation 2), and that they are identical as

$$\left(1 - \frac{1}{2N}\right) \frac{1}{2N}(1 - \mu)^4 \tag{135}$$

note the power of 4 is because our two alleles have to have failed to mutate through 2 meioses each.

More generally the probability that our alleles coalesce in generation $t+1$ and are identical due to no mutation to either allele in the subsequent generations is

$$P(\text{coal. in t+1 \& no mutations}) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^t (1 - \mu)^{2(t+1)} \tag{136}$$

assuming that $\frac{1}{2N} \ll$ and $\mu \ll 1$ then we can approximate this as

$$P(\text{coal. in t+1 \& no mutations}) \approx \frac{1}{2N}e^{-t/(2N)}e^{-2\mu(t+1)} \tag{137}$$

to make this slightly easier on ourselves lets further assume that $t \approx t + 1$ and so rewrite this as

$$\approx \frac{1}{2N}e^{-t(2\mu+1/(2N))} \tag{138}$$

If we sample two alleles at random from the population we will not in general know when they share a common ancestor. In which case we will need to integrate out over when this coalescent event occured. Doing this we find the probability that our two alleles are identical due to no mutation on either ancestral lineage since the pair shared a common ancestor to be

$$\frac{1}{2N} \int_0^\infty e^{-t(2\mu+1/(2N))}dt = \frac{1/(2N)}{1/(2N) + 2\mu} = \frac{1}{1 + 4N\mu} \tag{139}$$

The probability that our our pair of alleles are non-identical is simply one minus this, i.e.

$$\frac{4N\mu}{1 + 4N\mu} \tag{140}$$

This compound parameter $4N\mu$, the population-scaled mutation rate, will come up a number of times so we'll give it its own name

$$\theta = 4N\mu \tag{141}$$

So all else being equal, species with larger population sizes should have proportionally higher level of polymorphism at neutral sites.

Pairwise Coalescent time distribution and the number of pairwise differences.
In the preceding calculation you'll note that we could first specify what generation a pair of sequences coalesce in, and then calculate some properties of heterozygosity based on that. That's because neutral mutations do not affect the probability that an individual transmits that allele, so don't affect the way in which we can trace ancestral lineages back.

As such it will often be helpful to consider the time to the common ancestor of a pair of sequences, and then think of the impact of that on patterns of diversity. The probability that a pair of alleles have failed to coalesce in t generations and then coalesce in the $t + 1$ generation back is

$$\frac{1}{2N}(1 - \frac{1}{2N})^t \approx \frac{1}{2N}e^{-t/(2N)} \tag{142}$$

thus the coalescent time of a pair of sequences (T_2) is approximately exponentially distributed with a rate $1/(2N)$. We'll denote that by saying that $T_2 \sim Exp\left(1/(2N)\right)$. The mean coalescent time of a pair of a pair of alleles is $2N$ generations

Conditional on a pair of alleles coalescing t generations ago there are $2t$ generations in which a mutation could occur. Thus the probability of our pair of alleles are separated by j mutations since they last shared a common ancestor is

$$P(j|T_2 = t) = \binom{2t}{j}\mu^j(1 - \mu)^{2t-j} \tag{143}$$

i.e. mutations happen in j generations, and do not happen in $2t - j$ generations (with $\binom{2t}{j}$ ways this can possibly happen). Assuming that $\mu \ll 1$, and that $2t - j \approx 2t$ then we can approximate the probability that we have j mutations as a Poisson distribution

$$P(j|T_2 = t) = \frac{(2\mu t)^j e^{-2\mu t}}{j!} \tag{144}$$

i.e. a Poisson with mean $2\mu t$.

As our expected coalescent time is $2N$ generations, the expected number of mutations separating two alleles drawn at random from the population is

$$\mathbb{E}(j) = 4N\mu = \theta \tag{145}$$

We'll assume that mutations never happen at the same site twice, i.e. no multiple hits, such that we get to see all of the mutation events that separate our pair of sequences (we'll call this the infinitely-many-sites assumption, which should be fine if $N\mu_{BP} \ll 1$). Thus the number of mutations between a pair of sites is the observed number of differences between a pair of sequences.

We'll denote the observed number of pairwise differences at putatively neutral sites separating a pair of sequences as π (we usually average this over a number of pairs of sequences for a region). So we can estimate of θ from π, $\widehat{\theta}_{\pi}$ by setting $\widehat{\theta}_{\pi} = \pi$. If we have an independent estimate of μ, then from setting $\pi = \widehat{\theta}_{\pi} = 4N\mu$ we can obtain an estimate of the population size N that is consistent with our levels of neutral polymorphism.

6.4 The effective population size.

In practice populations rarely conform to our assumptions of being constant in size with low variance in reproduction success. Real populations experience dramatic fluctuations in size, and there is often high variance in reproductive success. Thus rates of drift in natural populations are often a lot higher than the census population size would imply.

To cope with this population geneticists often invoke the concept of an effective population size (N_e). In many situations (but not all), departures from model assumptions can be captured by substituting N_e for N.

Specifically the effective population size (N_e) is the population size that would result in the same rate of drift in an idealized constant population size, obeying our modeling assumptions, as that observed in our true population.

If population sizes vary rapidly in size, we can (if certain conditions are met) replace our population size by the harmonic mean population size. Consider a diploid population of variable size, whose size is N_t t generations into the past. The probability our pairs of alleles have not coalesced by the generation t^{th} is given by

$$\prod_{i=1}^{t} \left(1 - \frac{1}{2N_t}\right) \tag{146}$$

note that this is simply collapses to our original expression $\left(1 - \frac{1}{2N}\right)^t$ if N_i is constant. If $1/(N_i)$ is small, then we can approximate $1 - \frac{1}{2N_i}$ by $\exp(-\frac{1}{2N_i})$. Such that if N_i is never too small

$$\prod_{i=1}^{t} \left(1 - \frac{1}{2N_i}\right) \approx \prod_{i=1}^{t} \exp\left(-\frac{1}{2N_i}\right) = \exp\left(-\sum_{i=1}^{t} \frac{1}{2N_i}\right). \tag{147}$$

In our constant population size case the probability of failing to coalesce is $\exp(-t/(2N))$. So the variable population coalescent probabilities are still of the same form but the exponent has changed. Comparing the exponent in the two cases we see

$$\frac{t}{2N} = \sum_{i=1}^{t} \frac{1}{2N_i} \tag{148}$$

so that if we want a constant effective population size (N_e) that has the same coalescent probability as our variable population we need to set $N = N_e$ and rearrange this to see

$$N_e = \frac{1}{\frac{1}{t}\sum_{i=1}^{t}\frac{1}{N_i}}. \tag{149}$$

this is the harmonic mean of the varying population size. Thus our effective population size, the size of an idealized constant population which matches the rate of genetic drift, is the harmonic mean true population size over time. The harmonic mean is very storngly affected by small values, such that if our population size is one million 99% of the time but drops to a 1000 every hundred or so generations, N_e will be much closer to 1000 than a million.

Variance in reproductive success will also affect our effective population size. Even if our population has a large constant size of N individuals, if only small proportion of them get to reproduce then the rate of drift will reflect this much small number of reproducing individuals. If only N_M males get to contribute to the next generation and N_F females get to contribute to the next generation. When our two alleles pick an ancestor, 25% of the time our alleles were both in a female ancestor in which case they coalesce with probability $1/(2N_F)$, and 25% of the time they are both in a male ancestor in which case they coalesce with probability $1/(2N_M)$. The remaining 50% of the time our ancestral lineages are in two individuals are different sexes in a generation so cannot coalescence. Therefore, our probability of coalescence in the preceding generation is

$$\frac{1}{4}\frac{1}{2N_M} + \frac{1}{4}\frac{1}{2N_F} = \frac{1}{8}\frac{N_F + M_M}{N_F N_M} \tag{150}$$

i.e. the rate of coalescence is the harmonic mean of the two sexes population sizes, equating this to $\frac{1}{2N_e}$ we find

$$N_e = \frac{4N_F N_M}{N_F + N_M} \tag{151}$$

Thus if reproductive success is very skewed in one sex (e.g. $N_M \ll N/2$) our effective population size will be much reduced as a result.

6.5   The coalescent process of a sample of alleles.

Usually we are not just interested pairs of alleles, or the average pairwise diversity, we are interested in the properties of diversity in samples of a number of alleles drawn from the

population. To allow for this instead of just following a pair of lineages back until they coalesce, we can follow the history of a sample of alleles back through the population.

Consider first sampling three alleles at random from the population. The probability that all three alleles choose exactly the same ancestral allele one generation back is $1/(2N)^2$. If N is reasonably large then this is a very small probability. As such it is very unlikely that our three alleles coalesce at once, a in a moment we'll see that it is safe to ingnore such unlikely events.

The probability that a specific pair of alleles find a common ancestor in the preceding generation is still $1/(2N)$. There are three possible pairs of alleles so the probability that no pair finds a common ancestor is

$$\left(1 - \frac{1}{(2N)}\right)^3 \approx \left(1 - \frac{3}{2N}\right) \tag{152}$$

in making this approximation we are multiplying out the right hand-side and ignoring terms of $1/N^2$ and higher.

More generally when we sample i alleles there are $\binom{i}{2}$ pairs, i.e. $i(i-1)2$ pairs, thus the probability that no pair of alleles coalesces in the preceding generation is

$$\left(1 - \frac{1}{(2N)}\right)^{\binom{i}{2}} \approx \left(1 - \frac{\binom{i}{2}}{2N}\right) \tag{153}$$

while the probability of any pair coalescing is $\approx \frac{\binom{i}{2}}{2N}$.

We can ignore the possibility of more than pairs of alleles (e.g. tripletons) simultaneously coalescing at once as terms of $1/N^2$ and higher can be ignored as they are vanishingly rare. Obviously there are in reasonable sample sizes there are many more triples ($\binom{i}{3}$), and higher order combinations, than pairs ($\binom{i}{2}$) but if $i \ll N$ then we are safe to ignore these terms.

When there are i alleles the probability that we wait until the $t + 1$ generation before any pair of alleles coalesce is

$$\frac{\binom{i}{2}}{2N}\left(1 - \frac{\binom{i}{2}}{2N}\right)^{t-1} \approx \frac{\binom{i}{2}}{2N}\exp\left(-\frac{\binom{i}{2}}{2N}t\right) \tag{154}$$

thus the waiting time T_i to the first coalescent event in a sample of i alleles is exponentially distributed with rate $\frac{\binom{i}{2}}{2N}$, i.e. $T_i \sim Exp\left(\frac{\binom{i}{2}}{2N}\right)$. The mean waiting time till any of pair within our sample to coalesce is $2N/\binom{i}{2}$.

When a pair of alleles first find a common ancestral allele some number of generations back further into the past we only have to keep track of that common ancestral allele for the

pair. Thus when a pair of alleles in our sample of i alleles coalesce, we then switch to having to follow $i - 1$ alleles back. Then when a pair of these $i - 1$ alleles coalesce, we then have to follow $i - 2$ alleles back. This process continues until we coalesce back to a sample of two, and from there to a single most recent common ancestor (MRCA).

To simulate a coalescent genealogy at a locus for a sample of n alleles we therefore simply follow this algorithm

1. set $i = n$.

2. We simulate a random variable to be the time t_i to the next coalescent event from
$$t_i \sim Exp\left(\frac{\binom{i}{2}}{2N}\right)$$

3. choose a pair of alleles to coalesce at random from all possible pairs.

4. set $i = i - 1$

5. continue looping of steps 1-3 until $i = 1$ i.e. the most recent common ancestor of the sample is found.

by following this algorithm we are generating realizations of the genealogy of our sample.

We will first consider the time to the most recent common ancestor of the entire sample (T_{MRCA}). This is

$$T_{MRCA} = \sum_{i=n}^{2} T_i \tag{155}$$

generations back. As our coalescent times for different i are independent, the expected time to the most recent common ancestor is

$$\mathbb{E}(T_{MRCA}) = \sum_{i=n}^{2} \mathbb{E}(T_i) = \sum_{i=n}^{2} 2N/\binom{i}{2} \tag{156}$$

using the fact that $\frac{1}{i(i-1)} = \frac{1}{i-1} - \frac{1}{i}$ with a bit of rearrangement we can rewrite this is

$$\mathbb{E}(T_{MRCA}) = 4N\left(1 - \frac{1}{n}\right) \tag{157}$$

so the average T_{MRCA} scales linearly with population size. Interestingly, as we move to larger and larger samples (i.e. $n \gg 1$) the average time to the most recent common ancestor is converging on $4N$. What's happening here is that in large samples our lineages typically coalesce rapidly at the start and very soon coalesce down to a much smaller number of lineages.

Above we argued that a mutation is only becomes a fixed difference if it is lucky enough to be the ancestor of the entire population. As we saw above this occurs with probability

$1/(2N)$. How long does is take on average for such an allele to fix within our population. We've just seen that it takes $4N$ generations for a large sample of alleles to all trace their ancestry back to a single most recent common ancestor. Thus it must take roughly $4N$ generations for a neutral allele present in a single copy within the the population to the ancestor of all alleles within our population. This argument can be made more precise, but in general we would still find that it takes $\approx 4N$ generations for a neutral allele to go from its introduction to fixation with the population.

The total amount of time in the genealogy (T_{tot})

$$T_{tot} = \sum_{i=n}^{2} iT_i \tag{158}$$

as when there are i lineages each contributes a time T_i to the total time. Taking the expectation of the total time in the genealogy

$$\mathbb{E}(T_{tot}) = \sum_{i=n}^{2} i\frac{2N}{\binom{i}{2}} = \sum_{i=n}^{2} \frac{4N}{i-1} = \sum_{i=n-1}^{1} \frac{4N}{i} \tag{159}$$

so our expected total amount of time in the genealogy scales linearly with our population size. Our expected total amount of time is also increasing with sample size but is doing so very slowly. To see this more carefully we can see that for large n

$$\mathbb{E}(T_{tot}) = \sum_{i=n-1}^{1} \frac{4N}{i} \approx 4N \int_{1}^{n} \frac{1}{i} di = 4N \log(n-1) \tag{160}$$

here we are approximating our sum by an integral, which will work for large n. So our expected total amount of time in the genealogy is growing with n but it is doing so very slowly. This again follows from the fact that in large samples the initial coalescence usually happens very rapidly, so that extra samples adds little to the total amount of time in the tree.

We saw above that the number of mutational differences between a pair of alleles that coalescence T_2 generations ago was Poisson with a mean of $2\mu T_2$. A mutation that occurs on any branch of our genealogy will be cause a segregating polymorphism in the sample (making our infinitely-many-sites assumption). Thus if the total time in the genealogy is T_{tot} there is T_{tot} generations for mutations. So the total number of mutations segregating in our sample (S) is Poisson with mean μT_{tot}. Thus the expected number of segregting in history a sample of size n is

$$\mathbb{E}(S) = \mu\mathbb{E}(T_{tot}) = \sum_{i=n-1}^{1} \frac{4N\mu}{i} = \theta \sum_{i=n-1}^{1} \frac{1}{i} \tag{161}$$

Thus we can use this formula to derive another estimate of the population scaled mutation rate, by setting our observed number of segregating sites in a sample (S) equal to this

expectation. We'll call this estimator $\widehat{\theta}_W$

$$\widehat{\theta}_W = \frac{S}{\sum_{i=n-1}^{1} \frac{1}{i}} \tag{162}$$

this estimator was devised by Watterson, hence the W.

6.6 Deviations from the constant population model.

We've seen previously that changes in our population size can be captured by an effective population size. However, this will only be a useful measure if population sizes vary rapidly enough, that the harmonic mean effective population size over short time periods ($\ll N_e$) is representative of the effective population size averaged over longer time periods. If this is not the case there is no one effective population size, as we can not approximate our rate of drift by a single constant population. Furthermore, we've ignored the effect of population structure and selection which will violate our modeling assumptions.

We can hope to detect violations from our constant population size neutral model, by comparing aspects of our dataset to their expectations and distributions under our neutral model.

For example we have devised two estimates of θ, $\widehat{\theta}_\pi$ and $\widehat{\theta}_W$, using expectations of different aspects of our data (pairwise diversity and number of segregating sites respectively). Under our constant neutral model if we have sufficient data those two estimates should be equal to each other on average. But if there's some violation of our model they might not be. So one test statistic might be to take

$$D = \widehat{\theta}_\pi - \widehat{\theta}_W \tag{163}$$

which will be zero in expectation if our data was generated by a neutral constant population model.

6.7 The coalescent and population structure

Upto now we have assumed that our alleles that we have modelled in the coalescent setting are drawn from a randomly mating population such that any pair of lineages is equally likely to coalesce with each other. However, when there is population structure this assumption is violated.

We have previously written the measure of population structure F_{ST} as

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{164}$$

where H_S is the probability that two alleles sampled at random from a subpopulation differ, and H_T is the probability that two alleles sampled at random from a subpopulation differ.

We can relate these to our pairwise coalescent process, as H_S is simply the probability that two alleles sampled from our subpopulation mutate before they coalesce. If our pair of alleles drawn from our sub-population coalesce at time T_S then the probability that they are different is

$$1 - e^{-2\mu T_S} \approx 2\mu T_S \tag{165}$$

where this last approximation assumes that $\mu T_S \ll 1$. Thus averaging over the coalescent time of our pair of alleles

$$H_S \approx 2\mu \mathbb{E}(T_S) \tag{166}$$

and similarly $H_T \approx 2\mu \mathbb{E}(T_T)$. So we can write

$$F_{ST} \approx \frac{\mathbb{E}(T_T) - \mathbb{E}(T_S)}{\mathbb{E}(T_T)} \tag{167}$$

so F_{ST} is an expression of the decrease in pairwise coalescent times with a subpopulation compared to the total population. This also gives us a way to work out analytical expression for F_{ST} under simple models of population structure.

A simple population split model Imagine a population of constant size of N diploid individuals that τ generations in the past split into two daughter populations each of size N_e individuals, who do not subsequently exchange migrants. The in the current day we sample an equal number of alleles from both subpopulations.

The expected coalescent time of a pair of lineages sampled from the same subpopulation is simple $\mathbb{E}(T_S) = 2N_e$. The expected coalescent time when our alleles are sampled from different subpopulations is $\tau + 2N_e$, as our pair of lineages can not coalesce untill τ generations in the past. When we sample two alleles from our total population, 50% of the time they are drawn from the same subpopulation and 50% of the time they are drawn from different subpopulations so

$$\mathbb{E}(T_T) = \tfrac{1}{2}(\tau + 2N_e) + \tfrac{1}{2}2N_e = \frac{\tau}{2} + 2N_e \tag{168}$$

so that under this simple population split model

$$F_{ST} \approx \frac{\tau/2}{\tau/2 + 2N_e} = \frac{\tau/(4N_e)}{\tau/(4N_e) + 1} \tag{169}$$

i.e. F_{ST} at first linearly increases with the divergence time rescaled by the population size, such that the same divergence time causes less differentiation in larger populations.

Assuming that $\approx \frac{\tau}{2N_e} \ll 1$ then

$$F_{ST} \approx \frac{\tau}{4N_e} \tag{170}$$

A simple model of migration between an island and the mainland   We can also use the coalescent to think about patterns of differentiation under a simple model of migration drift equilibrium. Lets consider a small island population that is isolated from a large mainland population, and that both of these populations are constant in size. Our island has a population size N_I that is very small compared to our mainland population.

Each generation some low fraction m of our individuals on the island have migrant parents from the mainland the generation before. Our island may also send migrants back to the mainland, but these are a drop in the ocean compared to the large population size on the mainland and their effect can be ignored. The expected coalescent time for a pair of alleles sampled on the mainland is $\mathbb{E}(T_M)$

If we sample an allele on the island back and trace its ancestral lineage backward in time, each generation our ancestral allele have a low probability m of being descended from the mainland in the proceeding generation. So the probability that a pair of alleles sampled on the island are descended from a recent common ancestral allele on the island, is simply the probability that our pair of alleles coalesce before either lineage migrates. The probability that neither lineage migrates to the mainland in the preceding generation is $(1-m)^2 \approx e^{-2m}$, so the probability that neither of migrates to the mainland moving backwards in time, before the pair coalesce is

$$\int_0^\infty e^{-2mt} \frac{1}{2N_I} e^{t/(2N_I)} = \frac{1/(2N_I)}{1/(2N_I) + 2m} = \frac{1}{4N_I m + 1} \tag{171}$$

note the similarity of this argument to that by which we derived the expected heterozygosity (see work leading to eqn. (140)).

For a pair of alleles sampled on the island the time to either a coalescent event on the island or a migration event off the island is an exponential random variable $\sim Exp(1/(2N_I) + 2m)$, so our expected time back to our first event is

$$\frac{1}{1/(2N_I) + 2m} = \frac{2N_I}{1 + 4N_I m} \tag{172}$$

If this is a coalescent event, which it is with probability eqn. (171), then we are done. If this first event is a migration event of one of our lineages in the island (which it is with probability 1- eqn. (171)) we then have to wait on average $1/m$ generations to the other allele migrates to the mainland. Once both of our alleles are on the mainland they take a further $\mathbb{E}(T_M)$ generations on average until the coalesce. The expected coalescent time of a pair of alleles sampled on the island is

$$\mathbb{E}(T_I) = \frac{2N_I}{1 + 4N_I m} + \frac{1}{1 + 4N_I m} \times 0 + \frac{4N_I m}{1 + 4N_I m} \left(\frac{1}{m} + \mathbb{E}(T_M) \right) \tag{173}$$

This is a somewhat ugly expression, but if we assume that that the times $2N_I$ and $1/m$ are small compared to $\mathbb{E}(T_M)$ then

$$\mathbb{E}(T_I) \approx \frac{4N_I m}{1 + 4N_I m} \mathbb{E}(T_M) \tag{174}$$

this assumption is equivalent to assuming that the vast proportion of polymorphism on the island is due to the migration of different alleles from the mainland.

We'll assume that we have sampled alleles from the mainland and island roughly in proportion to their population sizes such that $\mathbb{E}(T_T) \approx \mathbb{E}(T_M)$, i.e. our heterozygosity in our total population is dominant, such that

$$F_{ST} \approx \frac{\mathbb{E}(T_M) - \mathbb{E}(T_I)}{\mathbb{E}(T_M)} \tag{175}$$

We can then substitute our expressions for $\mathbb{E}(T_T)$ and $\mathbb{E}(T_S) = \mathbb{E}(T_I)$ into F_{ST}. Doing so we arrive at

$$F_{ST} \approx \frac{1}{1 + 4N_I m} \tag{176}$$

7 The effect of linked selection on patterns of neutral diversity

A newly derived allele with an additive selection coefficient s will take a time $\tau = \approx 2\log(2N)/s$ generations to reach to fixation within our population. This short time window offers very little time for recombination between the the selected site and linked neutral sites.

First lets imagine examining variation at a locus fully linked to our selected locus, just after our sweep reached fixation. A pair of neutral alleles sampled at this locus must both trace their ancestral lineages back through to the neutral allele on whose background the selected allele initially arose. As that neutral allele, which existed τ generations ago is the ancestor of the entire population at this locus. Our individuals who carry the beneficial allele are, from the perspective of these two alleles, exactly like a rapidly expanding population. Therefore, our pair of neutral alleles sampled at our locus will be forced to coalesce $\approx \tau$ generations ago. This is a very short-time scale compared to the average neutral coalescent tie of $2N$ generations of a pair of alleles.

If we now allow recombination into our model we can think about a pair of alleles sampled at a neutral locus a recombination distance r away from our selected site. Our pair of alleles will be forced to coalesce $\approx \tau$ generations if neither of them reside on haplotypes that the selected allele recombined onto during the sweep. This is equivalent to saying that neither of our neutral alleles recombine off of the beneficial allele's background moving backward in time.

The probability that our lineage fail recombines off our beneficial allele's background and onto the ancestral background in the jth generation back is

$$r(1 - X(j)) \tag{177}$$

so the probability (p_{NR}) that our lineage fails to recombine off in the τ generations it takes our selected allele to move through the population is

$$p_{NR} = \prod_{j=1}^{\tau} \left(1 - r(1 - X(j))\right) \tag{178}$$

assuming that r is small then $(1 - r(1 - X(j))) \approx e^{-r(1-X(j))}$, such that

$$p_{NR} = \prod_{j=1}^{\tau} (1 - r(1 - X(j))) \approx \exp\left(-r\sum_{j=1}^{\tau} 1 - X(j)\right) = \exp\left(-r\tau 1 - \widehat{X}\right) \tag{179}$$

where $\widehat{X}$ is the average frequency of the derived allele across the trajectory $\widehat{X} = \frac{1}{\tau}\sum_{j=1}^{\tau} X(j)$. As our allele is additive its trajectory for frequencies < 0.5 is the mirror image of its trajectory for frequency > 0.5, therefore it average frequency $\widehat{X} = 0.5$. So

$$p_{NR} = e^{-r\tau/2}. \tag{180}$$

The probability that both of our lineages fail to recombine off the sweep and hence are forced to coalesce is p_{NR}^2, assuming that they coalesce at a time close to τ so that they recombine independently of each other for times $< \tau$.

If one or other of our lineages recombine off the sweep it will take them on average $\approx 2N$ generations to find a common ancestor as we are back our neutral coalescent. Thus the expected time till our pair of lineages find a common ancestor is

$$\mathbb{E}(T_2) = \tau \times p_{NR}^2 + (1 - p_{NR}^2)(\tau + 2N) \approx \left(1 - p_{NR}^2\right) 2N \tag{181}$$

where this last approximation assumes that $\tau \ll 2N$. So the expected pairwise diversity for neutral alleles at a recombination distance r away from the selected sweep (π_r) is

$$\mathbb{E}(\pi_r) = 2\mu \mathbb{E}(T_2) \approx \theta \left(1 - e^{-r\tau}\right) \tag{182}$$

so diversity increases as we move away from the selected site, slowly exponentially plauteuing to its neutral expectation $\theta = 4N\mu$.

To get a sense of the physical scale over which diversity is reduced consider a region where recombination occurs at a rate r_{BP} per base pair per generation, and our locus is ℓ base pairs away from the selected site $r = r_{BP}\ell$ (where $r_{BP}\ell \ll 1$ so we don't need to worry about more than one recombination event occurring per generation). Typical recombination rates are on the order of $r_{BP} = 10^{-8}$, in Figure 5 we show the reduction in diversity, given by eqn. (182), for two different selection coefficients.

For our expected diversity levels to recover to 50% of its neutral expectation $\mathbb{E}(\pi_r)/\theta = 0.5$, requires a physical distance ℓ^* such that $\log(0.5) = -r_{BP}\ell^*\tau$ as using our expression for τ then $\ell^* = \frac{s}{r_{BP}\log(4N)}$. The width of our trough of reduced diversity depends on s/r_{BP}, so else being equal we expect stronger sweeps or sweeps in regions of low recombination to have a larger hitchhiking effect. So that a selection coefficient of $s = 0.1\%$ would reduce diversity over 10's of kb, while a sweep of $s = 1\%$ would affect $\sim$100kb.

7.1 A simple recurrent model of selective sweeps

We sample a pair of neutral alleles at a locus a genetic distance r away from a locus where sweeps are initiated within the population at some very low rate ν per generation. The waiting time between sweeps at our locus is exponential $\sim Exp(\nu)$. Each sweep rapidly transits through the population in τ generations, such that each sweep is finished long before the next sweep $(\tau \ll 1/\nu)$.

As before our chance that our neutral lineage fails to recombine off the sweep is p_{NR}, such that the probability that our pair of lineages are forced to coalesce by a sweep $e^{-r\tau}$.
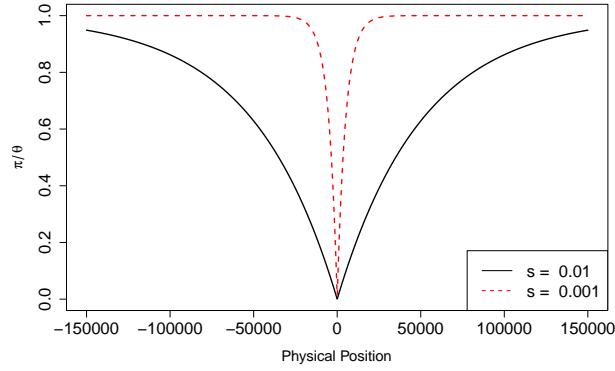
Figure 5: Reduction in diversity compared to its neutral expectation as a function of the distance away from a site where a selected allele has just gone to fixation. The recombination rate is $r_{BP} = 1 \times 10^{-8}$.

Our lineages therefore have a very low probability

$$\nu e^{-r\tau} \tag{183}$$

of being forced to coalesce by a sweep per generation. In addition of lineages can coalesce at a neutral rate of $1/(2N)$. Thus the average waiting time till a coalescent event between our neutral pair of lineages due to either a sweep or a neutral coalescent event is

$$\mathbb{E}(T_2) = \frac{1}{\nu e^{-r\tau} + 1/(2N)} \tag{184}$$

Now imagine that the sweeps don't occur at a fixed location with respect to our locus of interest, but now occur uniformly at random across our sequence. The sweeps are initiated at a very low rate of ν_{BP} per basepair per generation. The rate of coalescent due to sweeps at a locus ℓ basepairs away from our neutral loci is $\nu_{BP} e^{-r_{BP}\ell\tau}$. If our neutral locus is in the middle of a chromosome that stretches L basepairs in either direction the total rate of sweeps per generation that force our pair of lineages to coalesce is

$$2 \int_0^L \nu_{BP} e^{-r_{BP}\ell\tau} d\ell = \frac{\nu_{BP}}{r_{BP}\tau} \left(1 - e^{-r_{BP}\tau L} \right) \tag{185}$$

so that if L is very large ($r_{BP}\tau L \gg 1$) the rate of coalesce per generation due to sweeps is $\frac{2\nu_{BP}}{r_{BP}\tau}$. The total rate of coalescence for a pair of lineages per generation is then

$$\frac{2\nu_{BP}}{r_{BP}\tau} + \frac{1}{2N} \tag{186}$$

So our average time till a pair of lineages coalesce is

$$\mathbb{E}(T_2) = \frac{1}{\frac{2\nu_{BP}}{r_{BP}\tau} + \frac{1}{2N}} = \frac{r_{BP}2N}{\frac{4N\nu_{BP}}{\tau} + r_{BP}} \tag{187}$$
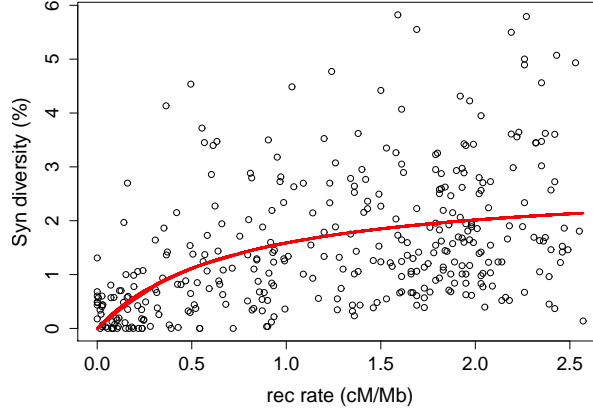
Figure 6: The relationship between (sex-averaged) recombination rate and synonymous site pairwise diversity (π) in *Drosophila melanogaster* using the data of Shapiro et al. 07 (kindly provided by Peter Andolfatto, see Sella et al. 09 for details). The curve is the predicted relationship between π and recombination rate obtained by fitting equation (188) to this data using non-linear least squares via the `nls()` function in R.

such that our expected pairwise diversity ($\pi = 2\mu\mathbb{E}(T_2)$) in a region of recombination rate r_{BP} that experiences sweeps at rate ν_{BP} is

$$\mathbb{E}(\pi) = \theta \frac{r_{BP}}{\frac{4N\nu_{BP}}{\tau} + r_{BP}} \tag{188}$$