# Outline of lectures 3 - 6

**Population genetics**

1. We want to know what theory says about the reproduction of genotypes in a population. This results in Hardy-Weinberg proportions. We imagine a population reproducing without any natural selection or any interference by any other forces such as mutation or migration.

2. This section is going to concentrate on what we expect to happen, based on an idealized simple population. Reality is of course messier, but to understand that reality, we first need to understand these simple situations.

3. The simplifying assumptions we make are that there is:

   - Random mating
   - No mutation
   - No migration
   - No differential viability
   - No differential fertility
   - Infinitely large population

4. If we have genotypes with current genotype frequencies $P$, $Q$, and $R$ of $AA$, $Aa$, and $aa$, and we count how many of the copies of the gene are $A$, they have a fraction $p = P + \frac{1}{2}Q$ of their genes being $A$ rather than $a$.

5. The rationale for this is simple: if a fraction $P$ of all individuals are $AA$, then of all copies of the gene in the population, a fraction $P$ of the gene copies that we find in gametes will be $A$ alleles that come from $AA$ parents.

6. Likewise a fraction $Q$ of all copies that we find in gametes will come from parent that are $Aa$ heterozygotes. Half of those will be $A$ copies.

7. The $p$ is the *gene frequency* (important – note the difference between this and a *genotype frequency*). The gene frequency of the $a$ allele is, for the same reasons, $q = \frac{1}{2}Q + R$.

8. The same numbers can also be computed by counting the fractions of $A$ and $a$ out of all copies of the gene found in the parent individuals.

9. Random mating is equivalent to random union of gametes. Imagine making a pot of all the female gametes that will be available to be contributed to the next generation, and a similar pot of male gametes. We could imagine drawing a pair of gametes, one from each, and making a diploid individual. That might be called random union of gametes.

10. This is equivalent to random mating because a random member of the offspring generation is descended from a random female and a random male, and Mendelian inheritance ensures that the gametes that each contributes to the next generation contain a random one of the two copies (at this locus) in that individual. So drawing a random parent, and then having it choose one of the two copies by Mendelian segregation, is equivalent to drawing one of the copies from the population of parents, at random.

11. The probability that the offspring gets a $A$ from the female parent is $p$, and (if male amnd female gene frequencies are equal) the probability that it gets a $A$ from the male parent is also $p$. Because these events are independent as a result of random mating, the probability that the individual is $AA$ is then $p^2$.

12. We can do the same for allele $a$. Suppose we call its gene frequency $q$. Since every allele is either $A$ or $a$, the frequencies $p$ and $q$ must sum up to 1.

13. The frequencies of the other genotypes $Aa$ and $aa$ can also be calculated by taking products of the gene frequencies.

14. The result is that $AA$, $Aa$, and $aa$ have expected genotype frequencies $p^2$, $2pq$ and $q^2$. (The 2 on the heterozygote genotype frequency is because there are two ways of making a heterozygote), $A + a$ and $a + A$.

15. The gene frequency in this offspring population is again $p$, since in that generation $P$ is $p^2$ and $Q$ is $2pq$, so that $p = P + \frac{1}{2}Q = p^2 + \frac{1}{2}(2pq) = p(p+q) = p$. (The $\frac{1}{2}$ cancels the 2 and a $p$ can be factored out, leaving the $(p+q)$). Note that $(p+q)$ is the sum of the frequencies of both alleles, hence it has to always be 1, so we can eliminate it from the product.

16. If we again mate these individuals randomly, the gene frequencies in the second generation are again $p$ and $q$.

17. Thus the genotype frequencies become these "Hardy-Weinberg" proportions, and stay that way forever. The gene frequencies remain forever $p$.

18. Mendelian genetic systems thus do not tend to lose genetic variability just because of random mating. The system of heredity that pre-Mendelian scientists such as Fleeming Jenkin used, blending inheritance, would lose it. The fundamental reason is that segregation in a heterozygote yields gametes that are $\frac{1}{2}A$ and $\frac{1}{2}a$, whereas in blending inheritance the genetic material itself gets averaged: it is as if they both become medium-sized A's.

19. The gene frequency of a sample of individuals can be computed in two different ways, both of which get the same answer.

    (a) We can take the numbers of the three genotypes ($n_{AA}, n_{Aa}$, and $n_{aa}$, and compute the genotype frequencies $P, Q$, and $R$ from them by dividing by their sum. The we get the frequency of $A$ among the gametes those individuals produce, which is $p = P + (1/2)Q$.

(b) We can simply count the numbers of $A$ and $a$ genes among the individuals in our sample, and get the gene frequency from that:

| Genotype | AA | Aa | aa |
|---|---|---|---|
| Number | $n_{AA}$ | $n_{Aa}$ | $n_{aa}$ |
| $A$ copies: | $2n_{AA}$ | $n_{Aa}$ | - |
| $a$ copies: | - | $n_{Aa}$ | $2n_{aa}$ |

The result is that the frequency of $A$ is $p = (2n_{AA} + n_{Aa})/(2N)$, where $N = n_{AA} + n_{Aa} + n_{aa}$ is the total number of individuals in our sample.

If you note that the genotype frequencies are $P = n_{AA}/N$, and $Q = n_{Aa}/N$, you will see that the two ways of computing the gene frequency are equivalent since $p = P + \frac{1}{2}Q$.

20. Here is an example. Suppose we find in a survey of a population two alleles, with the three genotypes $AA$, $Aa$, and $aa$ present in the sample in numbers 126, 98, and 45. That is a total of 269 diploid individuals so the genotype frequencies are, respectively $126/269 = 0.4684$, $98/269 = 0.3643$, and $45/269 = 0.1672$. The gene frequency (allele frequency) of $A$ is then $0.4684 + \frac{1}{2} \times 0.3643 = 0.6505$ (and for $a$ the allele frequency is 0.3495). We can alternatively simply count the total number of copies of $A$ in the sample, and divide by the total number of copies of the gene at this locus: $(126 \times 2 + 98)/(269 \times 2) = 0.6505$, and we can do the same for $a$.

21. In the above argument, we have used all of the six assumptions that were listed above. Exercise: Look at each one and try to see where it was used in the above argument.

22. When we relax the assumption of no differential viability and no differential fertility, we now have natural selection going on. It will no longer be true that gene frequencies remain unchanged. This is where Gregor Mendel is joined by Charles Darwin.

23. The *absolute fitness* of each genotype is the expected contribution a newborn individual of that genotype makes to the next generation. This is t he product of $\frac{1}{2}$(viability)(fertility). The one-half i s because each offspring it has only gets one-half of its genes from that parent.

24. Many populations (all, really) are subject to density-dependent population size regulation. If this were not true the earth would be totally covered with a constantly-rising layer of (say) chipmunks. If we make the simplest assumption, that this density regulation falls "fairly" on all genotypes, then it simply multiplies all viabilities by the same number, and/or multiplies all fertilities by the same number. It will do this if the density-dependent population size regulation acts at a different life stage than the natural selection that we are following, in a way unrelated to whatever causes the other fitness differences.

25. If this is true, then the ratios of the absolute fitnesses do not change as the population changes density, only the multiplier that makes them into abolute fitnesses. If we kill 10% of the chipmunks without regard to genotype, the numbers of chipmunks changes, but the proportions of the different genotypes will not change.

26. If we can assume that density-dependent population size regulation is falls equally on each genotype, then we can more or less ignore the sizes of the population in different generations, and just assume that the proportions of the different genotypes remains the same through that process.

27. This has the interesting effect of removing a lot of the ecological details from the argument, and just concentrating on the genetics.

28. Then we can define the *relative fitness* of a genotype as the ratio between its absolute fitness and the absolute fitness of some reference genotype. Thus relative fitneses might be $1 : 0.8 : 0.7$ for the three genotypes, for example, if we take $AA$ as the reference genotype. Density-dependent mortality (or reduction of fertility) might reduce the absolute fitnesses to $0.1 : 0.08 : 0.07$, but then they would still stand in the ratios of $1 : 0.8 : 0.7$.

29. As we shall see in a moment, the formulas for change of gene frequency under natural selection depend only on these ratios (and thus on the relative fitnesses). So we can forget about the issue of regulation of population density, and just deal with gene frequencies, genotype frequencies, and relative fitnesses.

30. With a one-locus two-allele case, one can compute the gene frequency after natural selection in the following argument.

31. The genotype frequencies at the beginning of the generation are of course $p^2 : 2pq : q^2$. When we count them by their contributions to the next generation (as a result of differential survival and fertility) they are in the proportions $p^2 w_{AA} : 2pq w_{Aa} : q^2 w_{aa}$.

32. These three numbers don't add to one, usually. So we can make them into frequencies by dividing by their sum. The sum is the *mean fitness* $\overline{w} = p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}$, a weighted average of the fitnesses, each weighted by its frequency in the newborns. Note that this is also the average value of the relative fitness of a randomly chosen newborn.

33. Dividing the sum by this quantity $\overline{w}$, we get the three frequencies:

$$\frac{p^2 w_{AA}}{\overline{w}} : \frac{2pq w_{Aa}}{\overline{w}} : \frac{q^2 w_{aa}}{\overline{w}}$$

Taking the frequencies of these three after selection (i.e. according to their contributions to the next generation) the gene frequency of $A$ in that next generation will be the frequency of $AA$ plus half that of $Aa$ or

$$p' = \frac{p^2 w_{AA}}{\overline{w}} + \frac{1}{2} \times \frac{2pq w_{Aa}}{\overline{w}} = \frac{p^2 w_{AA} + \frac{1}{2} \times 2pq w_{Aa}}{\overline{w}}$$

So we have ended up with the formula for the next generation's gene frequency, in terms of this generation's gene frequency and the fitnesses:

$$p' = \frac{p^2 w_{AA} + pq w_{Aa}}{p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}}$$

which by factoring $p$ out of the expression in the numerator, and calling the denominator $\overline{w}$:

$$p' \;=\; p\,\frac{(p\;w_{AA} + q\;w_{Aa})}{\overline{w}} \;=\; p\,\frac{\overline{w}_A}{\overline{w}}$$

34. In the denominator, the expression $p^2\,w_{AA} + 2p(1-p)\,w_{Aa} + (1-p)^2\,w_{aa}$ is just the weighted average of the $w$ for each genotype. It is the average answer if we choose individuals at random, from a population that is in Hardy-Weinberg proportions, and we ask each individual to report what its fitness is supposed to be, based on its genotype. It is called $\overline{w}$, the *mean fitness*.

35. The numerator term $(p\,w_{AA} + q\,w_{Aa})$ is a bit more mysterious, but it also has a simple interpretation. For an allele $A$, it is the average fitness of the individuals that this $A$ finds itself in, because $p$ of the time its partner allele is $A$, so that it is in a $AA$. and $q$ of the time it finds itself in a $Aa$.

36. Thus if we make a weighted average of the fitnesses in these two cases we get this quantity, which we denote as $\overline{w}_A$.

37. Note the rightmost expression: it says simply that the new gene. frequency is the old one ($p$) times the average fitness of the genotypes that a randomly-chosen $A$ allele happens to find itself in ($\overline{w}_A$), divided by the average fitness of everybody $\overline{w}$. In short, the gene frequency will increase if the mean fitness of $A$'s is bigger than the mean fitness of random individuals.

38. Since these formulas for the gene frequencies in natural selection have $w$'s in every term of the numerator and in every term of the denominator, anything that multiplies the fitnesses of all the genotypes by the same number does not affect the gene frequencies. This shows that the outcome of natural selection depends only on the relative fitnesses.

39. *A numerical example.* Let's start a generation with the numbers of $AA$, $Aa$, and $aa$ being (as before) 126 : 98 : 45. Suppose that the probabilities of survival of these three kinds of individuals is 1.0 : 0.8 : 0.7. For simplicity we are going to let all the differences that are due to genotype be in this probability of survival. Now suppose that each survivor contributes to one offspring in the next generation. We expect 126, 78.4, and 31.5 survivors, and thus also that number of contributions to the next generation. So there will be a total of 235.9 offspring contributed to the next generation. The gene frequency of $A$ in those contributions will be $(126 + \frac{1}{2} \times 78.4)/235.9 = 165.2/235.9 = 0.7002967$.

40. If, on the other hand, the population was at a low population density, so that each surviving individual contributes 5 offspring to the next generation, the numbers contributed will be $126 \times 5$, $78.4 \times 5$, and $31.5 \times 5$. These will be 630 : 392 : 157.5. In those individuals the gene frequency will be $(630 + \frac{1}{2} \times 392)/1179.5 = 826/1179.5 = 0.7002967$. Look familiar? It should, because all we have done that affects the calculation of the gene frequency is to have five times as many of all genotypes. So both the numerator and the denominator have been multiplied by 5, and thus the result turns out the same. So if the effects of population density fall on all genotypes equally, they do not change the gene frequency.

41. *Rate of change of gene frequencies as a result of natural selection.* Suppose that the fitnesses of $AA$, $Aa$, and $aa$ are $(1 + s)^2$ : $1 + s$ : $1$. This is a simple pattern of natural selection in which the fitnesses are multiplied by $1 + s$ for each additional $A$ allele placed in the genotype. It is a case that is of interest because the algebra of gene frequency change can be solved exactly.

42. $s$ is called a *selection coefficient*. When $s = 0$ there is no natural selection changing the gene frequencies. The curve of gene frequency change is a logistic curve (see projections). The time taken to change between any two gene frequencies is (approximately) inversely proportional to $s$. So starting at a low frequency of $A$, we end up with a high frequency, and the time we take to get from, say, 0.01 to 0.99, is about 10 times longer when $s = 0.001$ than when $s = 0.01$.

43. For the more complicated patterns of the natural selection that we are coming to, it helps to note that in our model scheme, the population starts each generation in Hardy-Weinberg proportions, though perhaps at a new gene frequency each time. If one allele (say $A$) is rare in the population, then it will occur mostly in heterozygotes. For example, if $p = 0.03$, the frequencies of the three genotypes $AA : Aa : aa$ will be 0.0009, 0.0582, and 0.9409. So if there are 1,000,000 newborns, we expect on average to have 900 $AA$'s, 58,200 $Aa$'s, and 940,900 $aa$'s.

44. If we count how many $A$ genes are in heterozygotes, there will then be 58,200 of those, but only 1,800 of the $A$'s will be in homozygotes (note that each homoygote has 2 $A$'s). Conclusion: *A rare allele occurs mostly in heterozygotes.*

45. Another, simpler way of seeing this is that if you are a copy of the $a$ allele "riding" in a sperm, on your way to an egg and to become part of a diploid fertilized egg, what fraction of your time will you end up with your partner being $A$? If $a$ is rare, that will happen most of the time.

46. *Overdominance.* A particularly interesting, though not very common, case of selection is overdominance, when the fitness of the heterozygote is higher than that of either homozygote. Natural selection will then bring the gene frequency toward an interior equilibrium, retaining both alleles. Having both alleles at nonnegligible frequencies is called a *polymorphism*. The exact equilibrium gene frequency depends on the fitnesses (in fact, if fitnesses are written as $1 - s$ : $1$ : $1 - t$ the equilibrium frequency of $A$ is $t/(s + t)$. If we plot fitness against gene frequency we get a quadratic curve. The peak of this curve turns out to be precisely at that equilibrium gene frequency.

47. We aren't going to do the mathematical derivation, but this movement of gene frequencies can be rationalized verbally in terms of the mean fitness of $A$ compared to the mean fitness of everybody. In an overdominant case, when $A$ is rare it is present mostly in heterozygotes. In that case $A$ copies have a higher mean fitness than $a$ (which, being common, are mostly located in $aa$ individuals. So $A$ increases when rare. When $A$ is common and $a$ is rare, the argument is reversed, with $a$ being mostly in heterozygotes and having the advantage.

48. *Underdominance.* When $s$ and $t$ are both negative we have underdominance, so that the heterozygote is the worst genotype. This is also uncommon, but it does happen with many chromosome alterations, if we follow their frequency in the population as if they were alleles. In underdominance, the gene frequency will move continually away from an interior equilibrium, which is now an unstable equilibrium. The gene frequency tends toward 1 or 0. Which one it goes to depends on which side of the unstable equilibrium it started from. Note that the outcome depends on the exact starting point. The plot of fitness against gene frequency is again a quadratic curve, but now it has a minimum at the unstable equilibrium. The stable equilibria are now 0 and 1.

49. *Selection and fitness.* In all of these cases the gene frequency changes so that the mean fitness either improves or remains the same – it never declines. In each case the population "climbs" the *adaptive surface* or *fitness surface* until it comes to rest at the top. Incidentally, this is true for constant relative fitnesses, and for any number of alleles. It is not perfectly true when fitness is controlled by multiple loci. It is also not perfectly true if the relative fitnesses depend on the gene frequencies. We have assumed that they do not. But in a lot of cases it is still true that there is a net gain of mean relative fitness from the beginning of the evolution of the gene frequencies to the end.

50. So is "all for the best in this best of all possible worlds?" (At least in terms of evolution resulting in optimal organisms). The case of underdominance makes the critical point – it shows that while evolution at a single locus (with constant relative fitnesses) results in improvement of the mean fitness, the population can sometimes come to rest on an equilibrium which is not the highest possible one. It depends on the starting point. A gene is evaluated by natural selection against the backgrounds in which it occurs, and that decides whether it will increase. If the fitnesses of $AA$, $Aa$, and $aa$ are $1.2 : 0.7 : 1$, then when $A$ is rare it is mostly occurring with $a$'s in heterozygotes, which have fitness 0.7. By comparison, the $a$'s are occurring in homozygotes which have fitness 1. So on average $a$ seems to be better and, correspondingly, copies of it survive and reproduce better. But in fact, $AA$ would actually be the best genotype. Natural selection is not making a global assessment of the effects of combining alleles, so it misses this and we end up with a population which is all $aa$.

51. Thus the opportunistic nature of natural selection causes us to climb the nearest peak on the adaptive surface, not the highest one. If one could always do the latter in evolution, would we be able to fly (unaided) at 500 miles per hour, swim to the depths of the ocean, and bore through solid rock, all the while composing brilliant sonatas? There does not seem to be any way to know, without a comprehensive understanding of the effects of all possible genotypes of humans.