GENOME 453 J. Felsenstein
Evolutionary Genetics Autumn, 2015

Outline of lectures 15-16

Genetic Variation and Neutrality

1. Until 1966, evolutionary geneticists had a limited range of genetic variation they could examine, and nothing like an unbiased sample of variation at the gene level:

   - They saw morphological variation from individual to individual in natural populations, but could not know how many loci were varying for any of the characters.
   - They could look at fitness variation uncovered by making whole chromosomes homozygous (for example, using experimental crosses in *Drosophila*) but they did not know exactly how many loci were contributing to the effects. They were by definition concentrating on variation chosen because of its large fitness effects, so they couldn't know what these effects were at typical loci.

2. Previous to that time there were two major theories about variation at the genetic level. H. J. Muller's *classical* view was that most loci would have a very common "wild-type" allele, and mutation would maintain, in a mutation/selection balance, low frequencies of mutant alleles that were deleterious. Theodosius Dobzhansky's *balancing selection* view was that most loci would have multiple alleles maintained by strong overdominance or frequency-dependent selection.

3. The evidence available was weak: evidence of how much fitness declined in, say, *Drosophila* when one made a whole chromosome homozygous at once by a clever system of crosses involving chromosome inversions. This was invented by Muller in the 1920s. However it could not show what was happening at individual loci. There was lots of argument, but no dazzling insight, about what was going on at individual genes.

4. In the early 1960's people started using the technique of protein electrophoresis to study variation at individual enzyme loci.

5. Gel electrophoresis was invented by Oliver Smithies in 1957 (he won the Nobel Prize 50 years later for unrelated work). It was related to earlier paper electrophoresis methods. A sample of blood is put in a gel made of potato starch or acrilamide, and subject to an electric current for a few hours. This is done under pH and temperature conditions that do not denature the proteins. The gel is then stained for the product (or the substrate) of one particular protein, and bands are seen, where the active enzyme protein is on the gel.

6. These bands show how far the protein of that enzyme has migrated through the gel. It is affected by both charge and conformation of its protein molecule. The method can detect a single amino acid substitution (though some are not detected) and, most importantly, does so in a way that has nothing to do with the fitness effect of the substitution.

7. The single-locus studies in the early 1960's often found variation at such loci, but there was no overall survey to see how typical this was. People tended to publish "lo and behold" papers showing that they had found that their one locus showed variation. It wasn't clear how many other people found no variation at the locus they studied and decided not to publish.

8. In 1966, Lewontin and Hubby and (independently) Harry Harris surveyed populations (respectively *Drosophila pseudoobscura* and humans) at multiple loci. Both projects found a lot of variation, which was a bit of a surprise.

9. The amount of variation is usually summarized in one of two kinds of statistic: heterozygosity or polymorphism:

   - *Polymorphism* is the fraction of loci at which the commonest allele is less than 95% in frequency (i.e. all the rarer alleles add up to more than 5%).
   - *Heterozygosity* is measured by taking the gene frequencies at each locus, and computing
     $$1 - (p_1^2 + p_2^2 + p_3^2 + \ldots + p_{10}^2)$$
     which is the predicted heterozygosity, as it is 1-(the homozygosity). This is then averaged across loci.

10. Typical values of heterozygosity seen by protein electrophoresis would be about 15% for invertebrates, about 7% for vertebrates. The variation of these typical values is big. For example, in amphibians one can find groups with heterozygosity values around 15%. These are for a variety of enzyme loci, measured by electrophoresis.

11. One issue is whether these loci may be regarded as typical loci. Another is what is happening at the large fraction of the genome that is not coding for amino acids. Susumu Ohno argued in 1972 that most of our genome is "junk DNA". Geneticists had long realized that there was far more DNA in a genome than could be accounted for by the number of protein loci. And the population geneticists among them had realized that if a substantial fraction of that DNA were maintained in conserved sequences, there would be too much "mutational load" (reduction in fitness by mutation). Still, it came as a shock to realize that natural selection could not be fine-tuning the contents of the genome.

12. (In 2012 the ENCODE consortium announced its conclusion that there is actually very little junk DNA in eukaryotic genomes. Basically no molecular evolutionist supports them in this, and a bit later in the course you will hear me rant and rave on this.)

13. *[Back to protein coding loci ...]* Polymorphism and heterozygosity, although both are measures of genetic variability, do not necessarily show concordant patterns when we compare natural populations. But they often give similar pictures.

14. Electrophoretically detectable enzyme polymorphism reveals only a small fraction of the variability. Perhaps 2/3 of all base substitutions in the DNA are not detectable, either because they don't change the amino acid, or because they change it to something that has a similar charge and a similar size of side-chain.

15. This issue began to be examined more directly once DNA sequencing was available. The alcohol dehydrogenase (ADH) locus in *Drosphila melanogaster* has two electrophoretic alleles (S and F). Marty Kreitman (1984) sequenced 11 different copies of the ADH gene (at the time, before PCR was invented, about all that could be done with the time and resources he had) which turned out to have a total of 43 mutations. One single site was responsible for the difference between the two electrophoretic alleles and was the only site that changed the protein sequence. All the rest were synonymous or noncoding substitutions.

16. Electrophoresis was a major technique until about 1980, when restriction site digests began to be used. They could detect part of the variation at the DNA level, namely those changes that caused particular restriction sites to disappear or appear. Variation at many sites could be detected, and statistical corrections made to infer from that how much variation there was at the DNA level.

17. DNA sequencing was introduced in the late 1970s, but only became cheap enough with the development of PCR in 1985. The development of SNP chips in the 1990s is another way to get a large amount of sequence data. Note that although complete genome sequencing is getting less expensive rapidly, for most purposes it is not necessary to get complete genomes to get a good idea of the amount of variation or the relationship between genomes in different species or different populations within a species.

18. The pattern of variation at the DNA level is similar to what is seen at the protein level – a lot of variation. If we take two copies of the same chromosome (say the two copies that an individual has) we will see one difference about every 1000 bases. This varies a bit (from about 1 in 500 to 1 in 1500) depending on the species. It is a slightly different figure from how often a SNP (Single Nucleotide Polymorphism) is found, since that requires that more than two copies be looked at and that the variation is not rare among them. In humans, SNPs are found about once every 1500 nucleotide sites.

19. Lewontin and Hubby had pointed out that there were a number of possible explanations for having this much variability:

    - Balancing selection. The alleles could be overdominant, or could be under frequency-dependent selection which favored rare alleles and prevented them from being lost.

- Deleterious mutation. But they pointed out that this does not work, as selection against a deleterious mutation would hold it down to a low frequency, and the variation detected had gene frequencies of more than one allele at high frequency.
- Neutral mutations. Alleles that do not differ in fitness are introduced by mutation, and float around, subject to genetic drift. Most of them ultimately get lost but some rise to fixation.

20. The observations immediately contradicted Muller's view, as they projected that hundreds to thousands of loci would be heterozygous in a typical individual. Later it was also realized that Dobzhansky's mechanism of strong balancing selection would not fit the observations either. With hundreds of loci becoming homozygous when a whole chromosome was made homozygous, the Balancing Selection view would predict far stronger inbreeding depression (reduction of fitness by inbreeding) than was actually observed.

21. Many surveys of patterns of variation find suggestive patterns, such as higher heterozygosity in invertebrates than in vertebrates, but do not settle the issue of whether the variation is maintained by selection. They also find that some categories of loci, such as enzymes in the glycolytic pathways, are less variable.

22. Some interesting cases, some of the first studies of "conservation genetics":

    - Horseshoe crabs (*Limulus polyphemus*) have shown a fair amount of stasis in phenotype over hundreds of millions of years. Do they lack variability at the genetic level? No, showed Robert Selander using electrophoresis in the 1970s. They have normal levels of variability.
    - The Northern Elephant Seal (*Mirounga angustirostris*) was reduced to just a few individuals about 1892, and has now recovered to about 150,000 individuals (and still increasing). The Southern Elephant Seal, a very similar species, was not reduced to a small number of individuals, though it was hunted. Sure enough, the Northern Elephant Seal shows very little genetic variability – most loci are fixed for one allele.

23. Lewontin and Hubby had suggested that the data could be explained by either balancing selection or by neutral mutation. Motoo Kimura, the greatest population geneticist of his era, advocated and greatly developed the latter position, using his formidable theoretical powers to greatly advance understanding of neutrality. His colleague Tomoko Ohta has argued for the importance of nearly-neutral mutations.

24. If neutral mutations are occurring at a locus at a rate μ per copy per generation, and each one is to a new allele (the "infinite isoalleles" model) and the effective population size is N_e, Crow and Kimura showed in 1964 that the expected amount of heterozygosity at the locus is $4N_e\mu/(4N_e\mu + 1)$. The derivation is explained in the lecture projection slides but I'll repeat it here.

25. Suppose a random pair of copies of our locus is chosen from a population. Each can come from any of the $2N$ possible copies, but they must be different copies. What is the probability that both came from the same parent copy? Each has its parent be one of the $2N$ possible copies, and they might come from the same copy. The probability of this is $1/(2N)$. So $1 - 1/(2N)$ is the probability that they came from different copies. If we consider the probability F that two random (distinct) copies have the same allele, if there were no mutation it would be for this generation F', where that is

$$F' = \left(\frac{1}{2N}\right) \times 1 + \left(1 - \frac{1}{2N}\right) F$$

where F is the same quantity in the parents' generation.

26. But there is mutation. If either copy is a new mutant, they cannot be the same allele. The probability that both are not new mutants (new in the most recent generation) is $(1-u)^2$. So the probability of being the same allele is

$$F' = (1-u)^2 \left[\left(\frac{1}{2N}\right) \times 1 + \left(1 - \frac{1}{2N}\right) F\right]$$

27. If F is at equilibrium (if there have been enough generations of mutation and genetic drift at the same value of N and u), we can set F' to F, and solve for F.

28. Using tedious high-school algebra this gives

$$F = \frac{(1-u)^2 \frac{1}{2N}}{1 - (1-u)^2 \left(1 - \frac{1}{2N}\right)}$$

29. Throwing out the very small terms in the numerator and in the denominator that have two factors of u or that have a product of a u and a $1/(2N)$, we get finally

$$F = \frac{1}{4Nu + 1}$$

which is the (approximate) probability of homozygosity of a random pair of copies. The probability of heterozygosity is $1 -$ (that), so it is $4Nu/(4Nu+1)$

30. (That is about the only real derivation we will do all quarter).

31. The alleles continually turn over, with no equilibrium gene frequencies of any allele, but the level of variation is roughly predictable. (The effective population size is the population size corrected for other details of the life cycle that affect the rate of genetic drift).

32. Low selection coefficients can maintain alleles segregating in populations. All that is required is that $4N_e s > 1$ which means that, for $N_e = 10^6$, the selection coefficient s can be as low as 0.00000025 and still maintain the alleles.

33. Recall that, for protein electrophoretic variation, Lewontin and Hubby observed about 15% heterozygosity in *Drosophila*, Harris observed about 7% in humans. A $4N_e\mu$ value of about 0.18 which would be obtained by having $N_e = 10^6$ and $\mu = 1.8 \times 10^{-7}$ will do this.

34. Less variability at some loci or in some parts of the genome is compatible with both theories, as the neutral mutation theory says that the variation is not maintained by selection, but it does not rule out there being selection against deleterious mutants. It is thus *not* a statement that all mutations are neutral – "purifying selection" can remove deleterious mutants without invalidating the neutral theory.

35. The 1000 Genomes Project (known as 1KGP), which can be found at `www.1000genomes.org`, has shown that levels of variation are consistent with this generalization:

    - The least variation (both SNPs and insertion/deletion events) is in coding sequences in the exons.
    - There is also reduction in variation in transcription factor binding sites upstream from the gene, and at microRNA binding sites in the 3' untranslated region (3'UTR) downstream from it.
    - There is a reduction in variation also in the 5' UTR upstream from the coding sequences.
    - There is some reduction in 5' sequences further upstream from the 5'UTR, and in 3' sequences further downstream from the 3'UTR.
    - ... and there is some reduction in introns, more in the first intron than in the others.

36. Laboratory experiments such as "population cages" with *Drosophila* can rule out large selection coefficients above 0.01, thus rejecting Dobzhansky's view, but they are totally incapable of detecting whether selection is 0 or (say) 0.0001, as lab experiments involve smaller populations and much shorter time spans than apply in natural populations. For example a population of 10,000 individuals (about the limit for a population cage), in a period of 100 generations (about the limit too, that's 8 years or so), could not detect any natural selection below about 0.0002 even if it were there.

37. A selection coefficient as small as $1/(4N_e)$ can be effective in nature, and that can be far smaller than anything we can detect in the lab. Nature can run an "experiment" millions of generations long with hundreds of millions of flies.

38. The controversy remains largely unresolved after 35 years, although it is most likely that much of our "junk" DNA accumulates mostly neutral mutations. For amino acid variations at protein loci the controversy is still unresolved.

39. Studies on pairs of whole human genomes find hardly any regions where there is evidence of strong balancing selection.

40. At the same time, genome comparisons of closely related organisms show signs of non-neutral patterns of substitution at protein-coding loci. and this is argued to come down against the neutral mutation theory.

41. Whatever its ultimate fate, the neutral theory plays a major role as the "null hypothesis" against which comparisons can be made.