

# Population Genetic Inference

*Gil McVean*

Department of Statistics, University of Oxford

January 22, 2002

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>What is population genetics?</b>                       | <b>2</b>  |
| 1.1      | What is inference? . . . . .                              | 3         |
| 1.2      | The history of population genetics . . . . .              | 4         |
| <b>2</b> | <b>How is variation inherited?</b>                        | <b>5</b>  |
| 2.1      | Darwin and Mendel . . . . .                               | 5         |
| 2.2      | The population genetics of continuous variation . . . . . | 7         |
| <b>3</b> | <b>What maintains variation?</b>                          | <b>9</b>  |
| 3.1      | The neo-Darwinian synthesis . . . . .                     | 9         |
| 3.2      | Kimura's non-neutral theory . . . . .                     | 11        |
| <b>4</b> | <b>How much variation?</b>                                | <b>12</b> |
| 4.1      | Serological methods . . . . .                             | 12        |
| 4.2      | Protein electrophoresis . . . . .                         | 12        |
| 4.3      | Statistics of genetic variation . . . . .                 | 13        |
| <b>5</b> | <b>Can selection maintain so much variation?</b>          | <b>14</b> |
| 5.1      | Genetic load arguments . . . . .                          | 14        |
| 5.2      | The neutral theory of molecular evolution . . . . .       | 16        |
| <b>6</b> | <b>How much variation at the DNA level?</b>               | <b>18</b> |
| 6.1      | Techniques for analysing variation . . . . .              | 18        |
| 6.2      | Statistics of DNA polymorphism . . . . .                  | 19        |
| <b>7</b> | <b>What use are neutral mutations?</b>                    | <b>20</b> |
| 7.1      | A genealogical view of population genetics . . . . .      | 21        |

# Good Questions in Population Genetics

## 1 What is population genetics?

Like so many branches of biology, what we think of today as population genetics would hardly be recognised by the founding fathers of the discipline. If you had been studying population genetics 80 years ago, you would probably have been mapping microscopic traits in *Drosophila* or developing efficient crossing schemes for agricultural breeding, 30 years ago you may have been analysing levels of protein polymorphism and population differentiation. Today, if you work in population genetics, you are more likely to be interested in using DNA sequence variation to map disease mutations in humans, or sites of adaptive evolution in viral genomes.

But of course there is a link between all three types of study: the genetics of variation. Broadly speaking, population genetics can be defined as the study of the genetical basis of naturally occurring variation, with the aim of describing and understanding the evolutionary forces that create variation within species and which lead to differences between species. For example, Slide 2 represents sequence level variation in a human gene called LPL, thought perhaps to play some role in hereditary heart disease (Nickerson et al., 1998). The types of questions we might want to ask of the data are

- Can we detect a link between sequence variation and a predisposition to heart disease?
- What does variation in this gene tell us about the history of humans?
- Can we detect the influence of natural selection on the recent history of the gene?

And in turn such questions raise other, more technical, but still critical issues to do with sampling;

- How many individuals should I sample?
- How much sequence should I collect from each?
- What is the best way to choose individuals from their geographic range in order to answer my question?

This course is about how to answer such questions; the design and analysis of population genetic experiments. How to turn information on the way in which the molecules of heredity vary between individuals into an understanding of the evolutionary forces that have shaped the history of species.

## **1.1 What is inference?**

Inference is the process of turning information from a sample into an understanding of the processes that generated the data. Methods of inference vary in the extent to which an explicit model of the underlying process is assumed, and the statistical approach used. If an explicit model is assumed, inference is called parametric, and if no model is assumed, then inference is nonparametric. Throughout this course I will concentrate on parametric inference; the idea that we can use explicit models of the evolutionary and genealogical processes as a framework for estimating important parameters and testing hypotheses for why genetic variation is distributed in the way we find it.

Although relatively recent in origin, coalescent theory (Kingman, 1982) has become the most widely used way of modelling the genealogical process of DNA samples from populations (more on these terms in lecture 2). For this reason it will be the starting point for our introduction to population genetics modelling, although it usually only gets a couple of pages in textbooks on population genetics. It is a remarkably intuitive and flexible way of thinking about how DNA sequences are related to each other by an underlying genealogical tree, and can readily accommodate interesting demographic processes, recombination and some forms of selection. You will learn much more about this over the coming course.

The course has three central themes; how to model genealogical processes, how to infer parameters of these models given a set of DNA sequences, and how to test whether the model you have used represents an adequate description of the data. Model testing is a critical part of population genetics, because showing that the assumed model is incorrect is the basis of making interesting discoveries about what is really going on. For example, most of the models we will come across assume that natural selection has not played an important role in shaping genetic diversity. If we can show that the data are not adequately described by a model without selection, and that the deviation is in the direction we might expect if there really were selection, we can be reasonably confident that selection has been important in the history of the gene. Of course, we can never be completely sure that our model of the evolutionary process is accurate. Our goal should be to make statements about how much faith we have in different models, and formulate new hypotheses to test

## **1.2 The history of population genetics**

If you could go back in time to pay a visit to the founding fathers of population genetics, they would probably be shocked by what you had to say. Apart from the obvious point that they would not have heard of DNA, the scientific consensus on the causes of genetic variation has experienced an enormous shift during the century of population genetics. Our starting point for analysing population genetic data is typically that most mutations at the DNA level are of no selective importance. But until about 30 years ago, the great obsession of population genetics was how selection and mutation could interact to maintain genetic variation in populations. This lecture aims to chart the central questions that have shaped a century of population genetics research, and to follow the events that led to such a major shift in view.

## 2 How is variation inherited?

### 2.1 Darwin and Mendel

The field of population genetics was created almost exactly 100 years ago, prompted by the rediscovery of Mendel's laws of inheritance. But to understand the importance of this discovery it is important to go back even further, to the experiments of Mendel, and of course, to Charles Darwin.

Although Mendel did not realise it, his discovery that certain traits of seed coat and colour in peas are inherited in a particulate manner was critical to the widespread acceptance of Darwin and Wallace's theory of evolution by natural selection. In its most simplified form, Darwin's theory consists of just three statements.

- Organisms produce too many offspring
- Heritable differences exist in traits influencing the adaptation of an organism to its environment
- Organisms that are better adapted have a higher chance of survival

The problem was, the way in which Darwin envisaged inheritance differences between organisms would be rapidly diluted through mating. In particular, he envisaged a form of blending inheritance in which offspring were intermediate between parental forms. Mendel's discovery that traits could be inherited in a discrete manner of course changed that view. Though it was not until de Vries, Correns, and von Tschermak-Seysenegg independently rediscovered both the phenomenon, and consequently Mendel's work, that this was acknowledged.

The idea that traits can be inherited in such a simple manner is extremely powerful. And following from de Vries and the others, many different traits showing such simple patterns of inheritance were rapidly described. For example in humans, the most well known examples are traits such as the ABO blood group and

albinism. However, while the discovery of particulate inheritance solved one problem, it created an even greater one.

The problem was that many geneticists, de Vries among them, came to understand the genetic nature of variation simply in terms of large, discrete differences. That is, the difference between round and wrinkly peas, or the difference between pink and white flowers. But the Darwinian view is one of gradualism; that there exists a continuum of variation, on which natural selection can act. De Vries was the first to use the term mutation; and by mutation he meant changes in genetic material that led to large differences in phenotype. On the other hand, in the early 1900s naturalists and systematists were developing a coherent view of evolution by natural selection that rested almost entirely upon the notion of small changes. The views of saltationists like Goldschmidt, with his “hopeful monsters” and empiricists such as Dobzhansky seemed to be almost entirely incompatible.

The solution is of course that the gradual, quantitative differences of the neo-Darwinians are in fact composed of the cumulative effects of many different loci, each behaving in a Mendelian, particulate fashion. Nilsson-Ehle (1909), working on pigmentation in wheat kernels, showed that the additive contribution of just a few loci (three in his case) could generate an apparently continuous distribution of phenotype. Likewise, Morgan and his colleagues, working in the Fly-room at Columbia, showed that patterns of inheritance of bristle number in *Drosophila* behave in a Mendelian fashion (Morgan et al., 1915). Similar results were found by Jennings in *Paramecium* (Jennings, 1929). Also important were the artificial selection experiments of Castle on quantitative traits in rats (Castle, 1903), which showed that selection acting on genes of small effect is effective. In short, the link had been made between Mendelian inheritance and Darwin’s theory of evolution by natural selection.

## 2.2 The population genetics of continuous variation

The first major contribution of theoretical population genetics to the understanding of natural variation arose from the discovery that Mendelian inheritance could underlie apparently continuous traits. In 1918, Fisher published a paper demonstrating how the phenotypic variation in a trait, and correlations between relatives, could be used to partition variation into genetic and environmental components, and also how the genetic component could be further partitioned into terms representing additive, dominant and epistatic contributions across loci (Fisher, 1918). This finding, along with earlier work on quantitative theories of inbreeding, had two important consequences. First, it naturally gave rise to a method for estimating the genetic contribution to variation for any quantitative trait. Second, it provided a means of predicting the effect of any artificial selection regime, as practiced by agriculturalists; and of course a framework within which to develop more efficient methods of breeding crops and animals with more desirable qualities and quantities.

Traits affected by multiple loci are called polygenic traits, but the term multifactorial is often used in order to emphasise the importance of environmental influences. Multifactorial traits can be further broken down into three types

1. Continuous traits: height, birth weight, milk yield
2. Meristic traits: bristle number in *Drosophila*
3. Discrete traits with continuous liability: polygenic disease, threshold traits

Fisher's results provided a means of directly estimating the contribution of genetics to variation in the phenotype, a factor which is termed heritability. There are two formulations of the term heritability, one known as narrow sense heritability, the other as broad sense heritability. Narrow-sense heritability is defined as the correlation between parents and offspring for some trait. For example, if we plot mid-parent value against offspring value, and fit a linear model. The linear coefficient  $b$ , in the

model  $y = bx + c$  is estimated by

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad (1)$$

and the relationship between  $b$  and heritability ( $h$ ) is given by  $b = h^2$ . What is the importance of this number? There are two ways this can be approached. First, it tells us something about what would happen were we to carry out artificial selection experiments, something that is of fundamental importance to agricultural breeding experiments. If only individuals with a trait value greater than some threshold are allowed to breed, if the difference in the mean values of the selected and entire population is  $S = \mu_s - \mu$ , then the selection response, defined as the difference in the mean of the offspring and parental populations,  $R = \mu' - \mu$  is given by  $R = h^2 S$ . For obvious reasons, another term for this form of heritability is realised heritability. The other way in which we can think about the importance of heritability, is that it tells us something about its genetic basis. For example, if we are interested in reducing the incidence of some disease, an estimate of the heritability would give us an indication of whether it is worth trying to find genes involved in the disease. In fact, it turns out that we can write the phenotypic variance in a trait as a sum of factors

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2 \quad (2)$$

where the terms on the left are, respectively, additive genetic variance, dominance effects, epistatic effects and environmental effects, then “narrow-sense” heritability is an estimate of

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \quad (3)$$

Another method for estimating heritabilities, at least in humans, is to use twins. Because twins share the same intra-uterine environment, the comparison of mono- and di-zygotic twins should theoretically give an estimate of the total genetic contribution to variation. This measure, which includes the first three components of variance, is called the “broad-sense” heritability. Specifically, we can estimate

heritability from the relationship

$$h^2 = \frac{r_M - r_D}{1 - r_D} \quad (4)$$

Slide 12 shows estimates for various traits in humans. The most notable feature is the very considerable variation in heritabilities between traits, for example fertility has very low heritability, while height and finger-print properties have very high heritabilities. Cognitive properties, such as IQ measures and behavioural distinctions such as extrovertism typically have a heritability of about 0.5. At the low end of the scale are features such as fertility, which have most likely been under strong selection. Finally, a note of caution should be added; the measurement of heritabilities in humans is notoriously inadequate because of unaccounted environmental correlations. For example, monozygotic twins tend to be treated more similarly than dizygotic twins. Any estimate of heritability in humans should be treated both with caution, and as an upper limit, particularly for behavioural traits.

A similar approach can be taken with threshold traits only rather than measure correlations, you measure concordance of a trait (proportion of comparisons with identical phenotype). Slide 13 shows concordance for a number of clinical traits in humans, and a corresponding estimate of the genetic component (Lewontin, 1982).

### **3 What maintains variation?**

#### **3.1 The neo-Darwinian synthesis**

The period of activity in evolutionary biology in the 1930s and 40s has come to be known as the neo-Darwinian synthesis. Researchers from the very different fields of systematics, palaeontology, cytology and genetics were all amassing evidence that Darwin's gradualist theory of evolution by natural selection was both theoretically and empirically feasible, and that evidence for its influence was everywhere in nature.

During this time, three figures from theoretical population genetics stand out; Ronald A. Fisher, J. B. S. Haldane and Sewall Wright. Together, they were

concerned with providing a mathematical description of the evolutionary process; how selection, acting to change allele frequencies at the microevolutionary level, could lead to macroevolutionary processes such as adaptation and speciation. Each, though, had their own emphasis. Fisher, in *The Genetical Theory of Natural Selection* (1930) laid out a comprehensive view of the power of natural selection (with a considerable discussion of eugenics); he was also the founding father of the statistical concept of likelihood and wrote extensively on experimental design. Haldane is best known for *The Causes of Evolution* (1932), but produced a very diverse body of work on problems relating to both the theory of selection and genetic inference, as well as a number of books that brought the ideas of evolution to a wider audience. On the other side of the Atlantic, Wright published his famous paper *Evolution in Mendelian populations* in 1931 (Wright, 1931). Wright's particular concern was with the way in which genetic drift can influence evolution.

To many people, Wright goes somewhat further than either Fisher or Haldane in providing a mathematical basis for the Darwinian synthesis. Critics of population genetics, such as Ernst Mayr, have argued that evolutionary biology is about three areas - adaptation, speciation and extinction - but that population genetics only considers the first. However Wright was utterly absorbed in the way in which chance differences between populations can lead to evolution. Although he does not explicitly deal with the subject, his work has greatly influenced the way in which we think about how populations can diverge from each other, eventually leading to speciation. His idea of an adaptive landscape (Wright, 1932) is one of the most persistent images in the field.

To the population geneticist of the time, there can have been little doubt that natural selection played the key role in maintaining variation in populations. Classic studies of conspicuous variation such as mimicry in butterflies and sickle-cell anaemia suggested that processes such as heterozygote advantage and frequency-dependent selection could maintain multiple phenotypes in a population. Moreover, the mathematical models showed that even immeasurably small selection

coefficients could potentially be sufficient for natural selection to act on, if the population size was large enough.

### **3.2 Kimura's non-neutral theory**

The fourth major figure in theoretical population genetics, Motoo Kimura, was originally a physicist. In the mid 1950s he starting publishing works which used diffusion theory methods, originally introduced into population genetics by Wright and Fisher, to study the fate of alleles in populations (Kimura, 1955). These papers involve some very difficult maths, though the key results are surprisingly neat. Kimura provided expressions for the rate of evolution and expected patterns of polymorphism for selected mutations. Although he is mainly remembered for the neutral theory of evolution (see below) his work laid the foundation for much of what we understand about the behaviour of selected mutations.

Kimura was the first theoretician I have discussed to have the benefit of knowing that genetic information is stored in DNA within chromosomes. Although DNA had been chemically identified as a component of cells in the late 1880s, it was not until the experiments of Avery and his colleagues in the early 1940s that its role in heredity was discovered. Even then these results were not widely accepted, and it was not until 1952 that Hershey and Chase provided incontrovertible evidence; just one year before the famous discovery of the structure of DNA by Crick and Watson. One of the most striking things about the development of theoretical population genetics was just how little it owes to an understanding of the mechanistic basis of gene function. Ernst Mayr once (1963) derided the field as nothing more than bean-bag genetics. But to a large extent, it is that level of abstraction (or simplicity) which makes population genetics so powerful a tool.

## **4 How much variation?**

### **4.1 Serological methods**

Population genetics had, until the mid 1960s, been largely concerned with phenotypic variation; coat colour, milk yield, butterfly wing-spots. Since then, there have been a number of technical innovations that have enabled researchers to detect and quantify variation at the molecular level. The results from such surveys were to start a whole new revolution in population genetics theory.

Prior to this point, few methods of detecting variation were available to the empirical researcher. The most important was the use of serological methods to analyse antigenic diversity in blood cells. The injection of blood cells into rabbits causes them to raise an immune response, such that when antibodies extracted from the animal's serum are mixed with blood cells of the same type, the blood cells coagulate and precipitate; some that can be visualised on a microscope slide. By this method, an amazing level of molecular diversity was revealed on both the red and white blood cells. Of course we all know about the ABO system, and most will know about the Rhesus system too. But there are over 50 different blood groups identified. Antigenic diversity on the white blood cells, is even more amazing, controlled by multiple HLA loci within the MHC cluster on chromosome 6. Each locus has many different allelic forms, and the frequency of the alleles varies considerably between populations. For example, the second most common allele at HLA-A in Europeans, which represents about 16%, is at a frequency of only 1% in Japanese (Cavalli-Szforza and Bodmer, 1971).

### **4.2 Protein electrophoresis**

However, few proteins can be assayed by serological methods. For this reason, the discovery of a technique called protein electrophoresis in the mid 60s was of enormous importance. Proteins are made of amino acids, some of which carry either a positive or a negative charge. In solution, proteins act as electrostatically

charged particles. So, if they are placed in a gel of starch agar or another polymer, and oppositely charged poles are placed at either end of the gel, they will tend to move to the pole of the opposite charge, and the rate at which they move is a function of their charge and size. Differences in amino acid composition can cause differences in the overall charge. Protein variants that migrate at different rates are known as allozymes, or isozymes. After a period of time, the position of the proteins in the gel can be visualised either by staining, or by making use of enzymatic properties of the molecule.

Protein electrophoresis is remarkably effective at detecting protein variation. It is estimated that 85-90% of all amino acid substitutions result in electrophoretically distinct molecules. And following the introduction of the technique into population genetics by Harris (in humans) (1966) and Lewontin and Hubby (in *Drosophila*) (1966), levels of protein variation were assayed in a wide range of organisms.

### **4.3 Statistics of genetic variation**

Before considering the results of these experiments, it is necessary to describe how genetic, or protein variation can be quantified. There are two simple measures which are widely used; polymorphism and heterozygosity. Polymorphism is the proportion of loci at which different alleles can be detected. It says nothing about how many alleles, or what frequency they are, just whether any differences can be detected. Heterozygosity at a locus is the proportion of individuals at which two distinct alleles can be detected, with the obvious caveat that heterozygosity can only be measured in diploid individuals. Why should we be interested in these two particular measures of variation? The answer is that these numbers are the key quantities in any theoretical population genetic understanding of the forces of variation, and under certain models have a very simple relationship to underlying parameters of mutation, selection and population size.

When protein electrophoresis was used to survey phylogenetically diverse taxa, from humans to plants, a remarkably high level of polymorphism was consistently

found. For example, in 30 species of mammals, with an average of 28 loci surveyed per species, about 1 in 5 loci are found to be polymorphic, and heterozygosity is about 5% Nevo (1978). Invertebrates appear to be slightly more polymorphic than vertebrates, for example, polymorphism in *Drosophila melanogaster* is about 0.5 and heterozygosity at about 0.15.

## **5 Can selection maintain so much variation?**

### **5.1 Genetic load arguments**

Why should such levels of polymorphism have been considered high when they were first described? The answer is that up until that point the most widespread belief among evolutionary biologists, was that majority of variation was maintained by balancing selection, one form of which is heterozygote advantage in which an individual with two different alleles is fitter than one with identical ones (as in sickle-cell anaemia). The finding that a large proportion of loci are polymorphic was problematic to this theory, because it meant that natural selection must be being incredibly efficient at balancing polymorphisms at many loci. For example, if 30% of proteins show allozyme variation and there are in the region of 50,000 proteins encoded for by the human genome, then natural selection must be maintaining polymorphism at about 15,000 loci.

Why should this be a problem? The reason is very simple; heterozygotes do not just produce heterozygous offspring. Mendelian segregation ensures that homozygous offspring are produced as well. So just by chance, we expect the number of heterozygous loci to vary considerably between individuals. And as a consequence we can expect fitness to vary considerably between individuals. This type of argument gives rise to the notion of genetic load, a concept that has historically been very important in the development of population genetics theory. Genetic load is defined as the difference in fitness between a population and its theoretical optimum. For example, in the case of heterozygote advantage, the most fit popu-

| Genotype | Fitness | Frequency   |
|----------|---------|-------------|
| AA       | $1 - s$ | $x^2$       |
| Aa       | 1       | $2x(1 - x)$ |
| aa       | $1 - s$ | $(1 - x)^2$ |

Table 1: Genotype frequencies and fitnesses under balancing selection.

lation would be heterozygous at every locus. The actual population cannot achieve this because of Mendelian segregation. For example, consider the genotype-fitness relationships in Table 1.

Ignoring drift in a finite population, the equilibrium frequency of each allele is 0.5, so the genetic load due to that locus is

$$L = \frac{w_{opt} - \bar{w}}{w_{opt}} = \frac{s}{2} \quad (5)$$

So, if there are 30,000 loci, each maintained by heterozygote advantage with a small selection coefficient of say 1%, then if fitnesses are multiplicative across loci, the genetic load due to segregation (called the segregation load) is such that the ratio of average fitness to best possible fitness is about zero ( $10^{-50}$ ). In other words, if absolute fitness is relevant to the survival of a species, then humans should be extinct.

Although striking, the relevance of a theoretical population that can never exist is not clear; in any finite population no individual is going to have all loci heterozygous. We can, however, ask a more directly relevant question; given Mendelian segregation, we expect variation between individuals in the number of heterozygous loci. How much variation in fitness should we expect? The answer is staggering. The expected number of heterozygous loci is about 3,000 in humans, but the variance is considerable, such that if you rank people by the number of loci for which they are heterozygous, the difference in number between the 99.5 and the 0.5 percentile is about 200 loci. If heterozygote superiority of about 1% is responsible for maintaining polymorphism, the difference in fitness between such individuals would be about 7.5 fold. In other words, if heterozygote advantage were really

were maintaining all polymorphism, we would expect much greater variance in reproductive success than is observed.

There is also a genetic load argument about the rate of substitution. Haldane (1957) showed that for a selected mutation to fix in a population, there have to be on average  $Nsx$  selective deaths each generation (where  $N$  is the population size,  $s$  is the selection coefficient and  $x$  is the allele frequency in that generation). So to fix a mutation with a 1% selective advantage requires about  $4.6N$  selective deaths, in addition to all the chance deaths that are nothing to do with selection. By way of example, if every of the 50,000 odd genes in humans have fixed one advantageous mutation since the split with chimpanzees and the population size has been on average about 10,000, just under 2.5 billion selective deaths must have occurred in the 5 million years or so since the species diverged. By the time you add up the total number of individuals that have lived during this period, selection accounts for pretty much all of them.

There are counter-arguments to genetic load paradoxes (for example frequency-dependent selection does not suffer the same load as heterozygote advantage and selection can eliminate many deleterious mutations at once), and currently genetic load arguments have little widespread interest or acceptance (Ewens, 2000). However, the important point is to understand that this type of thinking led people to question the belief that natural selection was responsible for maintaining all variation.

## **5.2 The neutral theory of molecular evolution**

Genetic load arguments, coupled with observations demonstrating the constancy of rates of molecular evolution, and the growing molecular genetics data showing how little of the eukaryotic genome is actually involved in protein encoding, led Kimura (1968) and King and Jukes (1969) to the conclusion that the majority of changes at the DNA level are of little or no functional consequence to the organism.

Today, it is hard to appreciate just how revolutionary this argument was. For

decades, evolutionary biologists had been amassing evidence about the incredible power of natural selection for creating adaptation. Yet here was the claim that the vast majority of change in the genes and proteins which make an organism are completely neutral. Naturally, there was much opposition to the ideas, but when data on the rate of evolution and levels of variation at the DNA level began to accumulate; the neutral theory had to be taken seriously.

The central features of the neutral theory are

- The majority of mutations are either strongly deleterious, or of no selective importance. Deleterious mutations are rapidly removed from the population, so most variation within species and differences between species are the result of neutral mutations.
- The rate of molecular evolution due to neutral mutations is identical to the neutral mutation rate;  $k = f_{neutral}\mu$ , where  $k$  is the rate of substitution,  $f_{neutral}$  is the proportion of all mutations that are neutral and  $\mu$  is the mutation rate.
- The level of polymorphism in a population is a function of the neutral mutation rate and the population size (actually effective population size - more on that next lecture).
- Polymorphisms are transient (on their way to loss or fixation) rather than balanced by selection

There have been some refinements to the neutral theory since first proposed. Most notably, the nearly-neutral and slightly deleterious mutation hypotheses of Tomoko Ohta (1995) have stimulated a resurgence of interest in natural selection. However, the consensus amongst population geneticists is that much of the variation at the DNA level is the result of effectively neutral mutations.

## 6 How much variation at the DNA level?

### 6.1 Techniques for analysing variation

Allozymes studies provided a fascinating glimpse into variation at the molecular level, but it was not until the advent of techniques for directly assessing DNA variability that the complete picture began to emerge. The first technique to be developed was the use of restriction fragment length polymorphisms (RFLPs). The genomes of many bacteria contain enzymes called restriction endonucleases which are thought to be used in defence against phages, and which cut DNA at specific sequence motifs. As with proteins, DNA is a charged molecule and will move through a gel down an electrostatic gradient at a rate which is a function of its size. It is possible to then transfer the gel onto a membrane (a Southern blot), denature the DNA, and probe it with radioactively labelled homologous sequence. By using a series of restriction enzymes it is possible to build up a map of the DNA sequence in terms of the relative locations of different restriction sites. A certain proportion of differences at the DNA level will affect restriction sites, which results in different patterns of restriction fragments.

The advent of the polymerase chain reaction (PCR) changed the face of genetics. And that includes population genetics. The two most important ways in which PCR has enabled the large-scale analysis of genetic variation, are first through the use of microsatellite markers, and second, through sequencing. Microsatellites are very short motifs, typically only 2-4 base pairs long, which occur in tandem repeats within genomes. Their replication appears to be highly unstable, such that the number of repeats changes at a much higher rate than single point mutations. The average rate for point mutations is about  $10^{-8}$  per generation. For microsatellites, it is about 1000 times higher, at  $10^{-5}$  per generation.

The first study of DNA sequence level variation through complete sequencing was carried out in 1983 by Marty Kreitman on the Alcohol dehydrogenase gene of *Drosophila melanogaster* (Kreitman, 1983). Full DNA sequencing, or rese-

quencing as it is also known, is clearly the only way of knowing the true extent of genetic variability, but until recently it was both very expensive and laborious. The twin developments of PCR sequencing and precision electrophoresis has made sequencing a much easier task, though it is still relatively expensive compared to using microsatellites or allozymes. A common strategy among groups wishing to analyse sequence level variation in many individuals is to carry out a SNP (single nucleotide polymorphism) survey. The project starts by fully resequencing a small panel of sequences to identify a set of polymorphisms which have only two alleles, and which are at intermediate frequencies (the SNPs). These sites can then be rapidly assayed for in a much larger target sample, which may consist of many hundreds of individuals. While SNP surveys are clearly very powerful, they also raise serious issues about how to analyse the data (Kuhner et al., 2000).

## **6.2 Statistics of DNA polymorphism**

Just as for protein polymorphisms, we can characterise variation at the DNA level by a series of statistics. The most commonly used statistics are the number of sites which are polymorphic in a sample, the average pairwise differences between sequences, and the number of distinct haplotypes.

What have studies of DNA level variation told us that we were not to expect from patterns of allozyme variation? Two observations are of particular importance. The first point is that different types of mutation at the DNA level have different levels of polymorphism. In particular, sites in coding regions at which some or all mutations have no effect on the amino acid encoded have higher levels of polymorphism than sites where mutations lead to amino acid changes. This observation is clearly evidence in favour of the neutral theory, because the less constrained sites have higher effective mutation rates. However, it is not true to say that all mutations that leave the amino acid unaltered are neutral. Levels of polymorphism in noncoding regions, such as 5' and 3' untranslated regions and introns, tend to be less in *Drosophila* than synonymous diversity. These patterns suggest

that subtle constraints act on non-coding regions, perhaps through their influence on gene regulation, or their role nucleic acid structure and stability.

The other striking pattern revealed from comparisons of DNA and allozyme variation is their lack of concordance. That is, while allozyme variation is remarkably constant across species, DNA level variation shows a much greater range. For example, *Drosophila* has about twice the allozyme heterozygosity of humans, but 10 times the DNA variation. What might cause this discrepancy? One possibility is that there really are subtle selective effects maintaining allozyme variation, but the issue is not resolved.

## **7 What use are neutral mutations?**

The conclusion that the majority of mutations segregating within populations and fixed between them are neutral has enormous implications for population genetics. If we are interested in understanding the relationship between genetic diversity and phenotypic variation, the task facing us is formidable. We neither understand the molecular basis of phenotype well enough to be able to predict changes in function from changes in gene sequence, nor can we hope to test each mutation segregating in the genome for its contribution to the phenotype of interest.

Yet in other ways, the conclusion of neutrality is a great benefit, because modelling the evolutionary behaviour of mutations is much easier without selection. Specifically, if mutations are neutral then the genealogical process (the historical process which defines ancestry for a set of chromosomes) is completely separate from the mutational process. And (most importantly) the genealogy contains all the information about the underlying evolutionary processes (apart from mutation). You can therefore think of mutations as providing a way of estimating the underlying genealogy, from which we can try to infer parameters of recombination, genetic drift, migration, and so on.

Furthermore, we can also model some forms of natural selection in a similar

manner, by assuming that we are not looking at the selected mutation itself, but to linked, neutral variation. Again, under these circumstances the genealogy contains all possible information about the evolutionary process, and we can use the distribution of linked neutral variation to tell us about the selection process. Very similar ideas can also be used to identify the genomic location of mutations associated with genetic disease. There will be much more on these topics throughout the course.

## **7.1 A genealogical view of population genetics**

The approach we will take to population genetic inference throughout this course is that there is an underlying genealogy that contains information about the key evolutionary parameters, and that we can use the distribution of mutations among individuals to infer aspects of the genealogy. However, in reality we can never know the genealogy for sure, and even if we did, we would still have some degree of uncertainty about the evolutionary process that generated it.

The statistical challenges for modern population genetic inference are consequently considerable. The range of possible outcomes for any given set of evolutionary parameters is enormous, and DNA sequence data is highly dimensional. This makes using all of the possible information in a data set a daunting challenge, but fortunately, we can often get a long way by concentrating on particular aspects, or summaries of data. Devising ways of using more information in population genetics inference is a major theme in current theoretical population genetics. Whether the questions relate to human history or the location of disease-associated mutations, computationally intensive methods of inference are rapidly developing areas of research.

## References

- Castle WE (1903). The laws of heredity of Galton and Mendel and some laws governing race improvement by selection. *Proc. Amer. Acad. Arts Sci.* 39:223–242.
- Cavalli-Szforza LL, Bodmer WF (1971). *The Genetics of Human Populations*. San Francisco: WH Freeman and Company.
- Ewens WJ (2000). The mathematical foundations of population genetics. In R. S. Singh and C. B. Kimbas (Eds.), *Evolutionary genetics*, pp. 24–40. Cambridge University Press.
- Fisher RA (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Phil. Trans. R. Soc. Edin.* 52:399–433.
- Fisher RA (1930). *The Genetical Theory of Natural Selection*. Oxford University Press.
- Haldane JBS (1932). *The Causes of Evolution*. New York: Harper and Row.
- Haldane JBS (1957). The cost of natural selection. *J. Genetics.* 55:511–524.
- Harris H (1966). Enzyme polymorphisms in man. *Proc R. Soc. Lond. B* 164:298–310.
- Jennings HS (1929). Genetics of the Protozoa. *Bibliogr. Genet.* 5:105–330.
- Kimura M (1955). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symp. Quant. Biol.* 20:33–53.
- Kimura M (1968). Evolutionary rate at the molecular level. *Nature* 217:624–626.
- King L, Jukes T (1969). Non-Darwinian evolution. *Science* 164:788–798.
- Kingman JFC (1982). The coalescent. *Stoch. Proc. Appl.* 13:235–248.

- Kreitman M (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304:412–417.
- Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000). Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156:439–447.
- Lewontin R (1982). *Human Diversity*. New York: Scientific American Publications.
- Lewontin RC, Hubby JL (1966). A molecular approach to the study of genic heterozygosity in natural populations II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudo-obscura*. *Genetics* 54:595–605.
- Mayr E (1963). *Animal Species and Evolution*. Harvard University Press.
- Morgan TH, Sturtevant AH, Muller HJ, Bridges CB (1915). *The mechanism of Mendelian heredity*. New York: H. Holt and Co.
- Nevo E (1978). Genetic variation in natural populations. Patterns and theory. *Theor. Pop. Biol.* 13:121–177.
- Nickerson DA, Taylor SL, Weiss KM, Clark A. G. and Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, et al (1998). DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics* 19:233–240.
- Nilsson-Ehle H (1909). *Kreuzungsuntersuchungen an Hafer und Weizen*. Lunds Universit. Arsskr. 5:1–122.
- Ohta T (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* 40:56–63.
- Wright S (1931). Evolution in Mendelian populations. *Genetics* 16:97–159.

Wright S (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. 6th Int. Cong. Genet. 1:356–366.