POPULATION GENETICS

Hamish G. Spencer
Associate Professor, Department of Zoology, University of Otago, Dunedin, New Zealand

Keywords: Genetic variation, natural selection, genetic drift, mutation, population structure, mathematical models, neutral theory, evolution, molecular biology

Contents

Summary

Population genetics concerns the genetic variation seen in natural and artificial populations. The various processes and properties that generate the observed patterns of variation – population structure, mating patterns, mutation, migration, genetic drift and natural selection – have been mathematically modeled by a number of authors, and the more important features are summarized below. By comparing the predictions of these models, the relative importance of the different population genetic processes can be deduced for a particular genetic system in some population of interest. The mathematical theories can also be used to infer features of the evolutionary history of groups of organisms.

1. Introduction

Population genetics is that part of biology concerning the application of genetic principles at the population level. These populations may be natural or artificial and the organism may be any living thing: humans, *Drosophila* fruit flies, the mustard *Arabidopsis thaliana*, and the eubacterium *Escherichia coli* are species commonly studied by population geneticists. Above all, population geneticists ask questions about genetic variation: how much variation is present within and among populations?, how is this variation structured?, what factors are important in this structuring?, what can the

pattern of variation reveal about the ecological and/or evolutionary history of the population or species? In turn, population genetics provides the theoretical underpinning for evolution, as well as having applications in a wide variety of human endeavors, from conservation and systematics to agriculture and forensics.

2. Basic Principles

2.1. Genetic Variation

Populations of plants, animals and humans all exhibit genetic variation, usually manifested as phenotypic variation. Of course, not all genetic differences lead to changes in the phenotype. Moreover, some phenotypic diversity is not genetically based, instead reflecting differences in the individuals' developmental environments. The amount of genetic variation in a population can be quantified in a number of ways. For each different allele, the proportion of genes that are of that type is known as the allele frequency. Allele frequencies, therefore, sum to exactly one. The proportion of heterozygotes observed in a population is known as the heterozygosity, H_O. Heterozygosity may then be averaged over several loci to estimate the proportion of the genome that is heterozygous in a randomly selected individual. The proportion of loci with more than one allele is also used as a measure of the level of variation. Almost all loci will have some variation present due to mutation, however, and such rare variation is not usually of interest. Hence, a locus is considered polymorphic if the commonest allele is at a frequency of less than 0.95 (and in some cases a less restrictive 0.99). In the pioneering electrophoretic study of a population of *Drosophila pseudoobscura* from Wildrose, California, Lewontin and Hubby examined 18 loci, 5 of which possessed more than one allele; the level of polymorphism was, therefore 28%. The mean level of heterozygosity over all 18 loci was found to be 11%.

2.2. Hardy Weinberg Principle

The starting point for studying the way genetic variation is distributed within a population is a straightforward application of the binomial theorem known as the Hardy Weinberg Principle. Consider a sexually reproducing population of monoecious, diploid organisms with separate generations. Then, the frequencies of two alleles, *A* and *a*, will remain unchanged indefinitely provided (i) there is no migration in or out of the population, (ii) no mutation occurs, (iii) there are no selective differences among the different phenotypes and (iv) the population is infinitely large. Moreover, if mating is at random with respect to the different genotypes and *A* and *a* are found at respective frequencies p and q (= $1 - p$), then, after a single generation, the respective frequencies of the three possible genotypes, *AA*, *Aa* and *aa*, are henceforth given by the terms of the binomial expansion of $(p + q)^2$, i.e., p^2, $2pq$ and q^2. The constancy of the genotype frequencies in fact implies that of the allele frequencies since, the frequency of the *A* allele in subsequent generations is simply the sum of the frequency of the *AA* genotypes and half the frequency of the heterozygotes:

$$p' = p^2 + \tfrac{1}{2}2pq = p\left(p+q\right) = p\,. \qquad (1)$$

At first glance, the assumptions of the Hardy Weinberg Principle seem absurdly restrictive. No population is infinite and in a large enough population, some mutation must be occurring every generation. But the Hardy Weinberg Principle is easily

generalized, for example, to cover loci with more than two alleles or dioecious organisms. Moreover, for a large range of biologically plausible situations, its predictions are robust to deviations from assumptions (i) – (iv). For example, with mutation rates in the typical range of 10^{-5} (i.e., one in every 100,000 alleles changes into the other), the effect is undetectable over hundreds of generations. The chief value of the Hardy Weinberg Principle is that it provides a conceptual null hypothesis: in the presence of migration, for instance, how are allele and genotype frequencies altered? Conversely, the maintenance of allele frequencies under its assumptions implies that evolutionary change requires that at least some of them are violated.

When the Hardy Weinberg Principle applies, the heterozygosity in a population is simply $2pq$. If there are several alleles present at frequencies p_1, p_2, …, p_n, then the heterozygosity is simply $1-\sum_{i=1}^{n} p_i^2$, i.e., one less the sum of the homozygote frequencies.

2.3. Non-random Mating

The random-mating assumption of the Hardy-Weinberg Principle is crucial in the calculation of genotype frequencies as the products of the appropriate allele frequencies. It can, however, be violated in two fundamentally different ways. First, mating may occur more (or less) often than expected between related individuals, a phenomenon known as inbreeding. Inbreeding is particularly common in plants, where many taxa either facultatively or obligatorily self-fertilize. Small populations are also subject to inbreeding because fewer mates are available. Second, mating may occur more or less often than expected between individuals who are phenotypically similar; this sort of pairing is called selective or assortative mating.

A direct consequence of an increased level of mating among relatives is an increase in homozygosity, often manifested as an increase in the incidence of genetic diseases. Indeed, a completely inbred population consists solely of homozygotes. We can quantify the degree of inbreeding using the inbreeding coefficient, F, the proportional decrease in heterozygosity from that expected under Hardy-Weinberg, H_E:

$$F = \frac{H_E - H_O}{H_E} \qquad (2)$$

Because the degree of inbreeding depends on the relatedness of individuals within mating pairs, Eq. (2) will apply to all different heterozygous types equally, and so can be rearranged to give the observed frequency of any heterozygote, A_iA_j:

$$2p_ip_j(1-F) \qquad (3)$$

Similarly, the proportion of observed A_iA_i homozygotes is

$$p_i^2(1-F) + p_iF \qquad (4)$$

Eqs. (3) and (4) suggest that a population subject to inbreeding may be viewed as consisting of two parts: a randomly mating, Hardy-Weinberg portion with a frequency of $1-F$, and a totally inbred portion with a frequency of F.

An alternative interpretation of *F* is as the probability that two randomly chosen genes at a single locus in a population are identical by descent, i.e., descendants of a single gene from some particular ancestral generation. An individual with two identical by descent genes is said to be autozygous and must also be inbred and, barring mutation, homozygous. Not all homozygotes are autozygous, however; like unmutated heterozygotes, they are said to be allozygous. Note that these definitions are all relative to some particular ancestral generation, usually chosen to be sufficiently recent that mutation can be ignored.

Inbreeding alters genotype frequencies but leaves allele frequencies unchanged. Selective mating, in contrast, usually changes both because different phenotypes have different numbers of offspring as a consequence of their choice of mates. Indeed, many models of selective mating are formally equivalent to those of fertility selection (see below), in which each type of mating pair has its own fitness (i.e., fertility). This fitness is the counterpart to the probability of that mating pair forming in the selective-mating model. Assortative mating, an historically important special case of selective mating, however, leads only to genotype-frequency change because all genotypes have the same mean number of offspring. These equal fitnesses may arise under positive assortative mating, when phenotypic classes preferentially mate among themselves. This situation is often modeled by partitioning each phenotypic class into those that breed only with the same phenotype and the remainder, which are added to a randomly mating pool containing the non-assorting individuals of other phenotypes. Clearly, this form of selective mating has a similar effect to inbreeding, decreasing heterozygosity while leaving allele frequencies constant. Negative assortative mating describes the situation in which individuals prefer to mate with unalike phenotypes, but most realistic scenarios lead to allele-frequency change and should be described as selective mating.

2.4. Mutation

Mutation is the process by which an allele is converted to another allele (also called a mutation or mutant allele) as an error in the usual genetic processes. These errors include the substitution of one DNA base pair for another and the insertion or deletion of one or more base pairs. The term mutation also applies to chromosomal changes, such as inversions and translocations; all such events also introduce heritable change into a population. In the simplest models deriving from the Hardy Weinberg Principle, an *A* allele is converted to an *a* or vice versa. In more sophisticated models, such as the infinite-alleles model, every mutant is novel.

Clearly, mutation must be a rare event and, indeed, most alleles mutate at very low rates, of the order of 10^{-4} to 10^{-8} mutation events per generation. At the level of DNA nucleotide pairs, the rate is approximately 10^{-9} per generation over a wide range of organisms. Mutation rates can be elevated by exposure to certain forms of electromagnetic radiation (e.g., X rays) and chemicals known as mutagens (e.g., colchicine and caffeine).

Mutation is the ultimate source of genetic variation in populations. Nevertheless, since mutation rates are so low, mutation by itself could have a significant effect on populations only over a very long time. To see this, consider a population with a frequency, *p*, of an allele, *A*, and subject to mutation at a rate, μ, per generation to another allele, *a*. In other respects, however, the population obeys the assumptions of

the Hardy Weinberg Principle. Since each generation a proportion, 1 - μ, of the *A* alleles do not mutate, the frequency of *A* after a single generation is given by

$$p' = (1-\mu)\, p \qquad (5)$$

and, thus, after *n* generations by

$$p^{(n)} = (1-\mu)^n\, p\,. \qquad (6)$$

If, for example, μ is 10^{-6} and the population starts fixed for *A* (i.e., $p = 1$) after 1000 generations the frequency of *A* has only reduced to $p^{(1000)} = 0.999$. In reality, other population genetic forces (especially selection and drift) act upon the effects of mutation to engender evolutionary change.

2.5. Migration and Population Structure

Allowing migration in and out of the study population also violates the Hardy Weinberg assumptions. Migration from an outside source can introduce novel variation in a similar way to mutation. Such immigration, however, can be very much more effective at changing a population's genetic makeup because so much more variation can be brought in by the immigrants each generation. In the same way as above, suppose the migrants are from a population fixed for the *a* allele and that the migrants make up a proportion, *m*, of the study population each generation. The frequency of the *A* allele after *n* generations of immigration is then given by

$$p^{(n)} = (1-m)^n\, p\,. \qquad (7)$$

If just 1 % of the population is made up of migrants, a population initially fixed for *A* will have only 90.4% *A* alleles after just 10 generations. Eventually, the *A* allele is swamped by this gene flow from the source and becomes extinct.

There are a number of different models that have a more sophisticated view of migration. Stepping stone models envisage a line (or sometimes a ring) of populations that exchange migrants only with their immediate neighbors; island models allow exchange between all populations. In all of these models migration each generation renders the populations more similar to each other, eventually completely homogenizing them.

Many models in population genetics presuppose that mating is at random throughout the population. One way this assumption can be violated is if there is geographical structure to the population, which can be modeled by envisaging it as collections of subpopulations or demes connected by some degree of migration (which may be zero). The degree of genetic differentiation among the demes may be quantified using what are known as Wright's *F* statistics or fixation indexes.

As a simple example, consider a population consisting of *k* equal-sized demes, and suppose that in each deme we know (or can estimate) the allele frequencies and hence their heterozygosities assuming Hardy-Weinberg proportions. Let H_S be the mean of

these within-deme heterozygosities and $\overline{p}_i$ be the average frequency of the *i*th allele over all *k* demes. The total heterozygosity (assuming Hardy-Weinberg proportions again) is then $H_{\mathrm{T}} = 1 - \sum_{i=1}^{n} \overline{p}_i^2$, which some simple algebra shows cannot be smaller than H_S. We then define the fixation index of the subpopulations relative to the total as

$$F_{\mathrm{ST}} = \frac{H_{\mathrm{T}} - H_{\mathrm{S}}}{H_{\mathrm{T}}}. \qquad (8)$$

Note the parallel with the inbreeding coefficient, *F*, in Eq. (2). Clearly, if all demes are genetically identical, $H_{\mathrm{S}} = H_{\mathrm{T}}$ and so $F_{\mathrm{ST}} = 0$, whereas if they are all fixed for different alleles $H_{\mathrm{S}} = 0$ and $F_{\mathrm{ST}} = 1$. F_{ST} values of less than 0.05 are usually interpreted as "little" differentiation, those between 0.05 and 0.15 as "moderate" and those from 0.15 to 0.25 as "great" and those above 0.25 as "very great." Values above 0.5 probably imply that separate biological species are confounded in the analysis.

This concept can be expanded to a hierarchy of levels: subpopulations within regions within the total species range, for example, giving F_{SR} and F_{RT}. Note that demes of different sizes and regions with different numbers of demes require that the averages are weighted by deme size and deme number, respectively, and, in practice, some adjustments need to be made for sampling effects at each level. We may still be interested in the overall level of differentiation of the subpopulations relative to the total, which can also be found from these other indexes using the relationship

$$1 - F_{\mathrm{ST}} = (1 - F_{\mathrm{SR}})(1 - F_{\mathrm{RT}}). \qquad (9)$$

This sort of approach can fruitfully show where within a species' range the greatest levels of differentiation occur.

One of the consequences of population subdivision revealed by Eq. (8) is a reduction in the proportion of heterozygotes in the total population compared with that expected under Hardy-Weinberg. Note that this deficit is observed even if the Hardy-Weinberg Principle applies exactly within each deme.

In the same way as an inbred population was partitioned above, a subdivided population may be viewed as consisting of two parts: a randomly mating, Hardy-Weinberg portion with a frequency of 1 - F_{ST}, and an inbred portion at a frequency of F_{ST}.

Indeed, the inbreeding coefficient, *F*, can also be incorporated into the hierarchical fixation-index taxonomy by viewing it as F_{IS}, the reduction in heterozygosity within a deme due to inbreeding. This approach also allows us to define an overall level of inbreeding, F_{IT}, which may either be calculated directly or from the analogue of Eq. (9).

A corollary of reduction in heterozygosity caused by population subdivision is an increase in heterozygosity when two previously isolated demes fuse, an effect known as the "Wahlund Principle." Plant and animal breeders are effectively using the Wahlund Principle when they outbreed their stocks to reduce the incidence of genetic diseases and deformities due to recessive alleles.

-
-
-

Bibliography

Bürger, R. (2000) *The Mathematical Theory of Selection, Recombination and Mutation.* Wiley, Chichester, UK. [A graduate-level text detailing recent work on mathematical models involving selection.]

Crow, J.F., and M. Kimura. (1970) *An Introduction to Population Genetics Theory*. Burgess, Minneapolis, Minnesota. [The standard introduction to mathematical models of population genetics.]

Ewens, W.J. (1979) *Mathematical Population Genetics*. Springer-Verlag, Berlin. [A graduate-level text on the mathematical aspects of the subject.]

Felsenstein, J. (1981) *Bibliography of Theoretical Population Genetics.* Dowden, Hutchinson and Ross, Stroudsburg, Pennsylvania. [Contains almost 8000 references to the primary literature up to 1980, searchable by author and words in the title.]

Fisher, R.A. (1999) *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Oxford University Press, Oxford, UK. [This facsimile of the 1st ed. Published in 1930 by Oxford's Clarendon Press was edited with a foreword and notes by J.H. Bennett. Although difficult to read, it was and remains highly influential.]

Hartl, D.L., and A.G. Clark. (1997) *Principles of Population Genetics*. 3rd ed. Sinauer, Sunderland, Massachusetts. [This textbook is one of the best and most wide-ranging introductions to the subject and is aimed at senior undergraduates or beginning graduate students. The mathematics required of the reader is usually fairly simple.]

Hedrick, P.W. (2000) *Genetics of Populations.* 2nd ed. Jones and Bartlett, Sudbury, Massachusetts. [Another excellent introductory textbook, using a slightly more mathematical approach than Hartl & Clark (1997).]

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK. [This book propounds the idea that genetic drift of selectively neutral alleles is the main agent of evolutionary change. Probably the most important book of the 1980s in the subject.]

Lewontin, R.C. (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York. [This book reviews the field of population genetics as it stood after the realization that genetic variation was ubiquitous, but no theory adequately explained this fact. Probably the most influential book of the 1970s in population genetics.]

Weir, B.S. (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer, Sunderland, Massachusetts. [An excellent description of statistical methods for analyzing a wide range of population genetic data, with numerous worked examples.]

Biographical Sketch

Hamish G. Spencer graduated with a BSc(Hons) in mathematics and an MSc in zoology from the University of Auckland, before obtaining a PhD from Harvard University, studying under the renowned population geneticist Richard C. Lewontin. In 1989 he was appointed as a lecturer in the Department of Mathematics and Statistics at the University of Waikato, and moved to the Department of Zoology at the University of Otago in 1992, where he is now an Associate Professor. He has written over 60 peer-reviewed papers for scientific journals and books, as well co-authoring one book and co-editing another. Most recently he has written entries for five different encyclopedia projects. His current research in mathematical population genetics looks at two areas: the maintenance of genetic variation by natural selection and the population genetic theory for genomically imprinted loci (at which the expression of a

gene depends on the sex of the parent from which it was inherited). He also writes about the history of genetics and the use of phylogenetics in evolutionary theory.

Selected Recent Articles:

Paul, D.B., and H.G. Spencer. 2001. Did eugenics rest on an elementary mistake? Pages 103-118 in *Thinking about Evolution: Historical, Philosophical and Political Perspectives*, R.S. Singh, C.B. Krimbas, D.B. Paul and J. Beatty (eds). Cambridge University Press, New York.

Spencer, H.G. 2000. Population genetics and evolution of genomic imprinting. *Annual Review of Genetics* 34: 457-477.

Kennedy, M., R.D. Gray and H.G. Spencer. 2000. The phylogenetic relationships of the shags and cormorants: Can sequence data resolve a disagreement between behavior and morphology? *Molecular Phylogenetics and Evolution* 17: 345-359.

Spencer, H.G., A.G. Clark and M.W. Feldman. 1999. Genetic conflicts and the evolutionary origin of genomic imprinting. *Trends in Ecology and Evolution* 14: 197-201.