Genetic and phenotypic variation

1 Phenotypic variation

Continuous variation refers to phenotypic differences that are measured on a quantitative scale. It is caused by two or more genes that slightly influence the trait, together with the influence of environmental factors. The distribution of a continuous trait in population is often close to a normal distribution, as a result of central limit theorem. The leading investigations on continuous variation are attributed to Francis Galton (1822-1911).

Discrete variation refers to phenotypic differences that can be assigned to a small number of clearly distinct classes. It is caused by segregation of alleles of a single gene. Environmental effects on discrete traits are small enough. Early investigations on the heritance of discrete variations are attributed to Gregor Mendel (1822-1884), whose paper was rediscovered in late 19th century.

2 Multifactorial inheritance

In natural populations, discrete traits are rare and the inheritance of most traits shows no clear-cut evidence for Mendelian segregation and the simple numberical ratios that Mendel observed in his pea-breeding experiments. This led to a great controversy in early 20th century. On the one hand were disciples of Galton, or biometricians, who dismissed the significance of Mendelian genetics. On the other hand were Mendelians, who argued that segregation of multiple genes could explain the normal distribution of continuous traits. The implications of multifactorial inheritance of discrete traits was the focus of 1918 paper by Ronald Aylmer Fisher (1890-1962).

The spirit of Fisher's model is that the value of a continuous trait is determined by multiple genes that segregate independently and influence the trait in an additive way. Let a continuous trait is determined by n genes and for ith gene, there are two alleles M_i and m_i. M_i adds 1 unit to the trait value while m_i is of no effect. Therefore, the contribution of each gene to the trait is a random variable that can be 2 (M_iM_i), 1 (M_im_i or m_iM_i) or 0 (m_im_i), and the trait value is the sum of n such variables. According to central limit theorem, distribution of the trait value is close to normal when n is large enough.

3 Maintaince of genetic variation

Darwin's theory of evolution assumes genetic variation among individuals. Therefore, early population genetics focused on the differences in genotype from one individual to another, and how such genetic variation is maintained from one generation to next, with emphasis on discrete traits. The results were intepreted to give support to either of two models on abundance and maintaince of genetic variation. The classic hypothesis asserted that genetic variation was uncommon and often harmful, and was maintained by a balance between recurrent harmful mutations and negative selection. The other one, balance hypothesis, declared that genetic variation was abundant and often good, and was maintained by selection favoring heterozygous genotypes or rare genotypes. Either classical or balance hypothesis overlooked the third alternative that genetic variation might have little or no influence on fittness, a model latter known as neutral theory.

4 Molecular population genetics

The issue of classic and balance hypothesis cannot be solved without unbiased methods enabling studying a large number of genes in different orginasms. This method, by looking into DNA, came at a price of disconnecting genotype from phenotype. It provides information about genotype, or DNA molecules, irrespective of its influence on phenotypes, due to complex interactions between genes and environment in the determination of physiology, development and behaviour.

Electrophoresis enables separating DNA or proteins. Protein electrophoresis is used primarily to study enzymes. The position to which a particular enzyme migrates is revealed by soaking the gel in a solution of substrates along with a dye that precipitates where the enzyme-catalyzed reaction takes place. If there is an amino acid replacement that changes the ionic charge of protein molecules, the alternative proteins will move to a different position. Such enzymes that differ in electrophoretic mobility as a result of allelic differences in a single gene is called allozymes. Hence, allozyme variation in a population is generally an indication of Mendelian variation.

Condiser a population containing an allozyme polymorphism with two alleles at different frequencies, the allele frequency refers to the proportion of all alleles of the gene that are of specific type. In population genetics, a polymorphism gene is one for which the most common allele has a frequency below 0.95. An allele is considered as a rare allele if its frequency is below 0.005. The 0.95 threshold for polymorphism gene and 0.005 for rare allele are arbitrary and serve as an attempt to focus on genes that have allele frequencies that are too high to be solely explained by recurrent mutations.

In a population, polymorphism can be measured by average proportion of polymorphism genes among genes. Heterozygosity can be measured by the average proportion of heterozygous genotypes among genotypes. Generally, there is a positive relationship between polymnorphism and Heterozygosity. Consider an ideal diploid population in which each new mutant allele is selectively neutral with negligible effect on survival and reproduction, the expected polymorphism P is given by

$$\ln(1-P) = \theta \ln(0.005) \approx -3\theta \tag{1}$$

$$\theta = 4N\mu \tag{2}$$

where N is population size and μ is mutation rate per gene per generation (Kimura Ohta, 1971). Under the same assumption, the expected heterozygosity H is given by

$$H = \frac{\theta}{1+\theta} \tag{3}$$

(Kimura Crow, 1964). Thus, the relationship between polymorphism P and heterozygosity H is

$$\ln(1-P) = \frac{-3H}{1-H} \tag{4}$$

Estimating polymorphism and heterozygosity is noisy. It may underestimate polymorphism since not all mutations can be detected by protein electrophoresis. It may overestimate polymorphism since it often only works for enzymes that are of high concentration in biological samples. However, protein electrophoresis still showed that polymnorphism is not as rare as suggested by classical hypothesis, which declares that most mutations are harmful and polymorphism is maintained by recurrent mutations. However, balance hypothesis assumes that polymorphism is maintained by selection favoring heterozygous or rare genotypes and thus suggests strong harmful effects of inbreeding, but observed inbreeding effects are often mild.

5 Polymorphism in DNA sequences

DNA electrophoresis separates DNA molecules by size. DNA fragments of a specific size can be produced by restriction enzyme, which cleaves DNA at specific site, or restriction site. Thus, digestion of genomic DNA with a restriction enzyme yields a set of fragments with different size according to the distances between restriction sites. These fragments can be separated by electrophoresis, rendered single-stranded by a solution of sodium hydroxide and transferred onto a filter. The filter is then bathed in a solution of labeled DNA probe, which binds to DNA fragments and show the positions. Thus, genetic variation of presence or absence of restriction sites can be identified as it changes the length of fragments from restriction enzyme digestion. A difference in the length of a restriction fragment found segregating in natural populations is called restriction fragment length polymorphism (RFLP).

Single nucleotide polymorphism (SNP) is present at a particular nucleotide site if DNA in the population frequently differ in the identity of nucleotide pair that occupies the site. In coding regions of genome, nucleotide polymorphisms that result in amino acid replacements are nonsynonymous polymorphisms, otherwise are synonymous polymorphisms. A unique combination of allelic states present along a single chromosome is a haplotype.

DNA sequencing data provides more comprehensive information on genetic variation. In a global multiple alignment of DNA sequences, segregating site refers to such nucleotide site that differs among other aligned sequences. Nucleotide mismatches are nucleotide sites that differ along any pair of aligned sequences.

Denote by S the number of segregating sites and by Π the average number of nucleotide mismatches along all pairwise alignments. A simple model that represents the relationship between S and Π is infinite-sites model. Assume that there are n aligned infinitely long DNA sequences with no recombination. Each nucleotide substitution occurs at different site and is selectively neutral. After sufficiently long time, the population reaches steady and S and Π are constants. At this state,

$$E(S) = \theta(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}) \tag{5}$$

(Watterson, 1975)

$$E(\Pi) = \theta \tag{6}$$

(Kimura, 1968). Here $\theta = 4n\mu$. μ is average rate of mutation per nucleotide multiplies number of nucleotides in each sequence being compared.