

STATISTICAL GENETICS

Ken G Dodds

AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand

Keywords: Quantitative Genetics, Relatedness, Inbreeding, Breeding, Heritability, Selection, Linkage, Mapping, Genetic markers, Quantitative trait locus

Contents

1. Introduction
2. Basic Principles
 - 2.1. Allele and Genotype Frequencies
 - 2.2. Hypothesis Tests
 - 2.2.1. Hardy-Weinberg Disequilibrium
 - 2.2.2. Linkage Equilibrium
 - 2.3. Segregation
3. Relatedness
 - 3.1. Inbreeding
 - 3.2. Kinship
 - 3.3. Estimating Relatedness
 - 3.4. Testing Relationships
 - 3.4.1. Exclusion
 - 3.4.2. Hypothesis Testing
4. Plant and Animal Breeding
 - 4.1. Infinitesimal Model
 - 4.2. Genetic Parameters
 - 4.3. Estimating Genetic Parameters
 - 4.3.1. Heritability
 - 4.3.2. Repeatability
 - 4.3.3. Maternal Effects and Dominance
 - 4.4. Selection
5. Locus Mapping
 - 5.1. Two-Point Linkage
 - 5.2. Multi-Point Linkage
 - 5.2.1. Ordering Loci
 - 5.2.2. Genetic Maps
 - 5.3. Physical Mapping
6. Quantitative Trait Locus Mapping
 - 6.1. Segregation Analysis
 - 6.2. Single Marker Analysis
 - 6.3. Interval Mapping
 - 6.4. Multi-Marker Methods
 - 6.5. Allele Sharing Methods
 - 6.6. Type I and II Errors
 - 6.6.1. Significance Thresholds
 - 6.6.2. Confidence Intervals for Location

6.6.3. Power

6.6.4. Designs to Increase Power

6.7. Fine Mapping

Glossary

Bibliography

Biographical Sketch

Summary

This paper gives an overview of the application of statistical methods to genetic data. The genetic data may take the form of allelic data from specific genes or from marker loci, or it may be phenotypic data from which we wish to infer something about the genetic components of the phenotypes. The types of analyses discussed here are of relevance to plant and animal breeding, as well as human genetics. The article covers the basic analysis of allelic data, relatedness, genetic improvement of plants and animals and the mapping of genetic loci and of quantitative trait loci. Related articles in this topic deal with the analysis of genetic data of populations (see *Population Genetics*) and the analysis of genetic data at the DNA sequence level.

1. Introduction

The scientific fields of statistics and genetics have developed side by side, with statistical analysis being applied to many types of genetic data, and with the field of genetics provoking new developments in statistical theory. In fact some modern parameter search procedures (“genetic algorithms”), which could be used in any field of statistical application, even rely on the principles of genetics. The first genetic principles were formulated by the Austrian monk, Gregor Mendel, in 1865. Although he did not apply statistical techniques (they had not been developed at that time), his data is amenable to such analysis, and subsequent researchers have done this. Mendel’s results were found to stand up to these analyses, except that there was evidence that the results fitted the hypotheses much better than would be expected by chance, and many authors have commented on possible reasons for this.

In the early 1900s, the English biometricians, Francis Galton and Karl Pearson, began applying the statistical techniques of correlation and regression to investigate the similarity of relatives. It was some years before these approaches were reconciled with the Mendelian principles, by R. A. Fisher and Sewell Wright. Since then until the late 1900s, most statistical analyses of genetic data have assumed that the trait of interest is controlled by either a few genes, or by the combination of many genes of small effect (the infinitesimal model). The former method was usually applied to discrete characteristics, while the latter method was usually (but not exclusively) applied to continuously measured (quantitative) characteristics.

Recently there has been widespread interest in finding genes that cause variation in quantitative characteristics, so called quantitative trait loci. The analyses of data to address this issue have required a combination of single gene and infinitesimal models.

Statistical analysis of genetic data is primarily concerned with the following areas of application: plant and animal breeding, medical genetics, forensic genetics, and the study of natural populations. Methods specific to the last of these are dealt with elsewhere (see *Population Genetics*).

2. Basic Principles

Some of the principles of statistical genetics underlie the methods used in all the application areas, and they will be dealt with in this section.

2.1. Allele and Genotype Frequencies

Suppose we are concerned with a single codominant locus, A , for a diploid organism, and that there are v possible alleles, A_1, A_2, \dots, A_v . The possible genotypes are the v homozygous types $A_1A_1, A_2A_2, \dots, A_vA_v$, and the $v(v-1)/2$ heterozygous types A_iA_j , $i < j$. The population is characterized by the proportions (or frequencies), P_{ij} , of these genotypes. These can be estimated by the corresponding proportions of these genotypes in a random sample of individuals from the population. Suppose we have a sample of size n , with n_{ij} individuals having genotype A_iA_j . Then

$$\hat{P}_{ij} = n_{ij} / n \quad (1)$$

is an intuitive estimator of P_{ij} . To find properties of these estimators we assume that the sample was drawn with a multinomial distribution. This is appropriate when the sample size is large (otherwise the hypergeometric distribution should be used). We find

$$\begin{aligned} \text{Var}(\hat{P}_{ij}) &= P_{ij}(1 - P_{ij}) / n \\ \text{Cov}(\hat{P}_{ij}, \hat{P}_{i'j'}) &= -P_{ij}P_{i'j'} / n, \quad (i, j) \neq (i', j'). \end{aligned} \quad (2)$$

Also of interest are the allele frequencies. These are estimated from the sample of genotypes as follows:

$$\text{Var}(\hat{p}_i) = (p_i + P_{ii} - 2p_i^2) / 2n, \quad (3)$$

where p_i is the frequency of allele A_i .

The variance still relies on the multinomial sampling properties for the genotypes, and is

$$\text{Var}(\hat{p}_i) = (p_i + P_{ii} - 2p_i^2) / 2n. \quad (4)$$

Equation (3) shows how allele frequencies can be derived from genotype frequencies, but to go in the reverse direction we need to make some assumptions about the genetic structure of the population. One of the simplest assumptions is that the population is in Hardy-Weinberg equilibrium, which requires that the population is large, undergoes random mating, and that there is no selection, mutation or migration. With this assumption

$$\begin{aligned} P_{ii} &= p_i^2 \\ P_{ij} &= 2p_i p_j, \quad i \neq j. \end{aligned} \tag{5}$$

The assumption means that alleles, as well as genotypes, are sampled at random from the population, and variances and covariances are found using the multinomial distribution for alleles, e.g.

$$\text{Var}(\hat{p}_i) = p_i(1 - p_i) / 2n. \tag{6}$$

Further complexity arises when loci are sex-linked, or when there are dominance relationships among the alleles (i.e. when one allele masks the presence or absence of another), when the population is sub structured or for other types (possibly mixed) of ploidy.

2.2. Hypothesis Tests

Genetic data may be collected with the aim of testing a particular hypothesis, or we may wish to test a hypothesis as a check on the validity of assumptions, before progressing with further analysis. This section discusses the tests for a number of hypotheses.

2.2.1. Hardy-Weinberg Disequilibrium

Hardy-Weinberg equilibrium can be tested using a chi-square goodness-of-fit test. For each genotype (i, j) we have the observed number (n_{ij}), and the expected number ($E_{ij} = n$ times the Hardy-Weinberg frequencies). We then calculate the chi-squared statistic:

$$X^2 = \frac{\sum_{\text{genotypes}} (O_{ij} - E_{ij})^2}{E_{ij}} \tag{7}$$

which is asymptotically distributed as χ_{df}^2 , where df, the degrees of freedom, is equal to $v(v-1)/2$, the number of genotypes minus the number of parameters estimated (one for each allele). There are a number of conditions for the validity of chi-square tests, such as all the E_{ij} being at least five. This is unlikely to be the case where there are many alleles, with several at low frequency, but there have been alternative tests developed, such as likelihood ratio tests, exact tests and permutation tests.

2.2.2. Linkage Equilibrium

When data are collected on more than one locus, one question we may wish to address is whether a pair of loci act independently, i.e. the probability of having a certain allele at one of the loci does not depend on the allele at the other locus. When this is the case the loci are in linkage equilibrium. Let a second locus be denoted by B , its alleles by B_j , $j=1,2,\dots,w$. Then the disequilibrium between alleles A_i and B_j is

$$D_{ij} = p_{ij} - p_i p_j \quad (8)$$

where p_{ij} denotes the gametic frequency of A_i/B_j combinations, p_i is the frequency of A_i and q_j is the frequency of B_j . To use the coefficient in this form requires that gametic data is sampled, or that gametic types can be inferred from family data, or (in the case of both loci on the same chromosome) that single chromosomes have been sampled.

Linkage equilibrium can be tested using a chi-square goodness of fit test, where the observed numbers are $n\hat{p}_{ij}$ (where n is the number of gametes sampled), while the expected numbers are $n\hat{p}_i\hat{q}_j$. The chi-square statistic can be written as

$$X^2 = \sum_{i=1}^v \sum_{j=1}^w \frac{n\hat{D}_{ij}^2}{\hat{p}_i\hat{q}_j} \quad (9)$$

where the disequilibrium coefficient is estimated (using maximum likelihood) by replacing frequencies with their estimates. The df for the test are $(v-1)(w-1)$. As in the case of testing Hardy-Weinberg equilibrium alternative testing methods are available and may be preferable when the number of alleles is large and/or the sample size is small.

When genotypes are scored, a direct count of A_i/B_j combinations is not usually possible. Under the assumption of random mating, genotypic frequencies are the products of gametic frequencies. This allows gametic frequencies to be estimated (for example, using the expectation-maximization algorithm). Without this assumption it will be necessary to estimate and test composite measures of genotypic disequilibria, incorporating both gametic and non-gametic (within individual) disequilibria.

2.3. Segregation

The experimental results that lead Mendel to propose a particulate model of inheritance were the observations of segregation patterns in his crosses. The data from experiments such as these can be tested to see whether they fit certain models, for example, by using a chi-square goodness of fit test. For example, in a cross where both parents have genotype A_1A_2 , we would expect progeny numbers in the ratios 1:2:1 for the genotypes A_1A_1 , A_1A_2 , and A_2A_2 respectively. If allele A_1 is dominant to A_2 , then genotypes A_1A_1 and A_1A_2 cannot be distinguished phenotypically. In this case progeny numbers are expected in the ratio 3:1 for $A_1A_1 + A_1A_2$ and A_2A_2 . Various other crosses can be tested in a similar way. Essentially these are testing that the alleles at the locus are segregating in the ratio 1:1, known as Mendelian segregation.

Mendel also performed experiments looking at more than one character at a time. One such experiment concerned seed color (green or yellow) and seed shape (round or wrinkled). In the F_2 generation (from pure breeding lines) he obtained the numbers shown in Table 1. It was hypothesized that each character had a dominant type (round dominant to wrinkled, yellow dominant to green), and that the two characters

segregated independently. In this case the types would be found in the ratios 9:3:3:1 as shown in Table 1. A chi-square goodness of fit test of these data gives $\chi^2 = 0.47$ with 3 df (not significant). Such a test is actually comprised of three components: a test of a 3:1 ratio for shape, a test of a 3:1 ratio for color, and a test of no association between the two characters. The χ^2 statistic can be partitioned into each of these components to give a 1 df test for each. The last of these (no association) is actually a test for linkage, a topic that will be covered in detail later because of its importance in medicine and agriculture.

Shape	Color	Number	Ratio
Round	Yellow	315	9
Wrinkled	Yellow	101	3
Round	Green	108	3
Wrinkled	Green	32	1

Table 1: Numbers of F₂ progeny for each combination of two characteristics

3. Relatedness

3.1. Inbreeding

Inbreeding arises when the parents of an individual are related, i.e. they have an ancestor in common. The inbreeding coefficient of an individual is defined as the probability of the two alleles at a locus being identical by descent (IBD), i.e. are copies of the same allele from a common (maternal and paternal) ancestor. If information on ancestors is known, then

$$F_X = \sum \left(\frac{1}{2}\right)^m (1 + F_A) \quad (10)$$

the inbreeding coefficient (F) can be calculated as where the sum is over all paths from one parent of the individual (X) to a common ancestor (A) and back to the other parent. The number of individuals in a path (not counting X) is denoted by m . As a simple example, suppose an individual's parents (M and P) are half-sibs, with common parent A . Then there is a single path relevant to the inbreeding calculation ($M \rightarrow A \rightarrow P$), and we find $F = (1 + F_A)/8 = 1/8$ if A is not inbred.

Inbreeding can also be estimated using genetic marker data, either for finding a population average inbreeding, or on an individual basis. To obtain good estimates, quite large genetic samples are required.

3.2. Kinship

A measure of the relationship between two individuals is the inbreeding of a hypothetical progeny of the two individuals, and is the probability that an allele chosen at random from one individual is IBD with an allele chosen at random from the other individual. This is called their coancestry, or coefficient of kinship (f). The relatedness

(r) is defined as twice this value. The genetic relationship between two individuals can be further delineated by considering the probability that they share 0, 1 or 2 alleles IBD at a locus. These probabilities are denoted by k_0 , k_1 and k_2 respectively, and they sum to one. In addition, $k_1^2 \geq 4k_0k_2$ for all valid relationships. We also have that

$$r = k_1 / 2 + k_2 \quad (11)$$

As shown in Table 2, some relationships, which have common values of r , have differing values of the k s.

Relationship	r	k_0	k_1	k_2
Self (Monozygous twin)	1	0	0	1
Parent-Offspring	0.5	0	1	0
Full-sibs	0.5	0.25	0.5	0.25
Half-sibs, Uncle-niece, Grandparent-grandchild	0.25	0.5	0.5	0
Cousin	0.125	0.75	0.25	0

Table 2: Relationship coefficients for some common relationships

3.3. Estimating Relatedness

There arise a number of situations in which the relatedness between pairs of individuals is desired. In the absence of known pedigree relationships, marker data can be used to infer this information. Relationship estimation can be used to avoid matings between close relatives, to estimate genetic parameters without requiring pedigreed populations, and to study social behavior (e.g. through reconstructing a genealogy).

There are a number of formulas suitable for estimating the relatedness between individuals. One of these is the measure of D. C. Queller and K. F. Goodnight:

$$\hat{r} = \frac{\sum_{i=1}^2 \sum_j \sum_l (p_{ijl} - \bar{p}_{jl}) \delta_{ijl}}{\sum_{i=1}^2 \sum_j \sum_l (p_{ijl} - \bar{p}_{jl}) \delta_{ijl}} \quad (12)$$

where i indexes the two individuals, j indexes loci, l indexes the 2 alleles of i , $i' \neq i$, i.e. refers to the other individual, and δ is an indicator with value 1 when i has allele l at locus j .

An alternative measure (M. Lynch and K. Ritland) is given by

$$\hat{r} = \frac{1}{W} \sum_j w_j \frac{p_a(S_{bc} + S_{bd}) + p_b(S_{ac} + S_{ad}) - 4p_a p_b}{(1 + S_{ab})(p_a + p_b) - 4p_a p_b} \quad (13)$$

where

$$w_j = \frac{(1 + S_{ab})(p_a + p_b) - 4p_a p_b}{2p_a p_b}, \quad W = \sum_j w_j,$$

j indexes loci, a and b are the two alleles for individual X and c and d are the two alleles for individual Y at the j th locus, S is an indicator function, taking the value one if the two alleles are the same (zero otherwise). The w are weights calculated as the inverse of the sampling variance assuming no relatedness. There are similar expressions for estimating the k coefficients.

An alternative to these methods is maximum likelihood estimation. The likelihood can be written as:

$$L = \prod_{loci} (k_0 P_0 + k_1 P_1 + k_2 P_2) \quad (14)$$

where for each locus, P_m is the probability of the observed genotypes conditional on m alleles being IBD. In general, numerical methods are required to maximize this likelihood and find the estimates of the k_m .

Accurate estimation of relatedness requires the use of many independently segregating loci. For example, using the Queller-Goodnight method to estimate the relatedness for full-sibs using 20 loci, each with three alleles at equal frequency, has a standard error of 0.13 (for the mean of 0.5).

-
-
-

TO ACCESS ALL THE 31 PAGES OF THIS CHAPTER,
Visit: <http://www.eolss.net/Eolss-sampleAllChapter.aspx>

Bibliography

Doerge R.W., Zeng Z-B. and Weir B.S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **12**, 195-219. [A discussion of statistical issues related to QTL mapping].

Falconer D.S. and Mackay T.F.C. (1996). *Introduction to Quantitative Genetics*, 4th edn, Longman. [An introductory level textbook for population and quantitative genetics].

Gianola D. and Hammond K. (Eds.) (1990). *Advances in Statistical Methods for Genetic Improvement of Livestock*, Heidelberg: Springer-Verlag. [A collection of 23 chapters from a symposium on the “state of the art”, as well as areas for further research, in livestock breeding methods].

Lange K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*, New York: Springer-Verlag. [Mathematical, statistical and computational principles for genetic analysis, concentrating on mapping].

Liu B.H. (1998). *Statistical Genomics – Linkage, Mapping and QTL Analysis*, New York: CRC Press. [Extensive coverage of the topics listed in the title, as well as their biological background].

Lynch M. and Walsh B. (1998). *Genetics and Analysis of Quantitative Traits*, Sunderland, MA: Sinauer. [This book gives detailed treatment to most of the topics covered here, as well as background genetic and statistical theory].

Mayo O. (1987). *The Theory of Plant Breeding, 2nd edition*, Oxford: Oxford University Press. [Covers the design and analysis of plant breeding programmes, with emphasis on traits and breeding systems that are important in plants].

Mrode R.A. (1996). *Linear Models for the Prediction of Animal Breeding Values*, Oxford, UK: CAB International. [A detailed treatment of the mathematical theory for the prediction of breeding values].

Ott J. (1999). *Analysis of Human Genetic Linkage, 3rd edition*, Baltimore, MD: Johns Hopkins University Press. [A comprehensive coverage of linkage analysis, oriented toward human data].

Weir B.S. (1996). *Genetics Data Analysis II*, Sunderland, MA: Sinauer. [This book covers topics in the analysis of discrete genetic data].

Biographical Sketch

Ken Dodds became interested in statistical methods for genetic data when growing up on a dairy farm in New Zealand. This interest continued while completing a BSc (Hons) at the University of Otago, New Zealand. Ken did his Ph.D. in the Statistics department at North Carolina State University, with a minor in Genetics. He studied under Bruce Weir, graduating in 1986. Since then he has worked for AgResearch at Invermay, near Dunedin, New Zealand, as a statistical geneticist involved in the analysis of animal breeding programs, and in the analysis of molecular genetic data for gene discovery, parentage testing and hybrid testing. He has spent 1-6 month sabbaticals at INRA – Toulouse (France); North Carolina State University (U.S.A.) and University of New England (Australia) and has tutored at the Summer Institute of Statistical Genetics at North Carolina State University. Ken is an Associate Editor of JABES, and has served as a committee member for the Association for the Advancement of Animal Breeding and Genetics.