Identification of major animal genes in field collected data by use of statistical methods. A review

T. Szwaczkowski

Department of Genetics and Animal Breeding
Poznań Agricultural University
Wałyńska 33; 60-637 Poznań, Poland

(Received 1 February 1993; accepted 21 June 1993)

ABSTRACT

The purpose of this paper is to present some of the methods currently available for the detection of major genes based on population data. These methods are concerned, primarily, with quantitative traits because their application to discrete and threshold traits is not always possible (e.g. tests of deviations from normality). As mentioned, only some methods for identification of the genes with large effects are presented. Directions of methodological investigations and their effects are shown. In the first part of this paper, existing statistical methods (adopted for major gene identification) are described. In the second part, methods of human genetics modified for animal genetics (Major Gene Index -- MGI) are presented. Finally, the method generally accepted currently as the most accurate (segregation analysis) is characterized. Some disadvantages connected with the application of this method are also described.

KEY WORDS: major genes, distribution of traits, segregation analysis

INTRODUCTION

The classical animal breeding theory for quantitative performance traits is based on a polygenic model of inheritance. However in recent years, several genes considerably affecting these traits have been identified. The major locus is defined as the one having an effect of at least one standard deviation of the metric traits, as measured by the difference between the two alternative homozygotes (Roberts and Smith, 1982). The Booroola gene affecting ovulation rate in sheep, the double muscling gene of cattle, the dwarf gene in poultry, the halothane sensitivity gene in pigs and the rapid postweaning growth gene in mice are notable examples. In Hanset's (1982) opinion several traits (e.g. twinning, calving ease, size and...
resistance to infections and parasitic diseases) are "candidates" for a mixed major gene and polygenic inheritance.

This situation requires modifications of breeding programmes, since mixed inheritance (genes with a large effect and genes with small effects) leads to an increase of genetic variability compared to polygenic inheritance of the traits. Hence, the effects of major locus often lead to an increase of heritability and genetic correlation. Major gene effect is usually masked by a large number of environmental factors, as well as additive and non-additive polygenic effects. Moreover, the ability to detect a major gene is a function of the magnitude of its effect. To make the best use of the variability induced by these major genes, it is necessary to identify the carrier genotypes and their frequencies, as well as the gene effects and activity in breeding stock. Most importantly, detection of a major locus is very important for gene mapping and for creating transgenic individuals. Different statistical techniques have been suggested for the identification of human, animal and plant genes with large effects (Karlin et al., 1979; Hoeschele, 1988; Elsen et al., 1988; Elsen and Le Roy, 1989; Thomas et al., 1991).

The aim of this study is to present some methods which may be applied to animal breeding field-collected data, mainly for continuous traits.

THE CLASSICAL STATISTICAL METHODS

General analysis of trait distribution

The polygenic inheritance of metric traits is connected with a normal distribution of their phenotypic values (e.g. Knott et al., 1990). Therefore, it is assumed that effects of major genes lead to deviations from normality. In the most spectacular case (indirect inheritance) frequency distribution of phenotypic values results in multimodality. However, a test of multimodality did not find wide application in detecting major locus because mixed inheritance (polygenic + major genes) is relatively rarely combined with more than one mode value. However it is not clear whether mixed inheritance is connected with the non-normality of phenotypic distribution. As measurements of deviations from normality two parameters: skewness ($\gamma_1$) and kurtosis ($\gamma_2$) have generally been used (Hammond and James, 1970; Hanset and Michaux, 1985). Testing of the hypothesis concerning non-skewness is based on the following test function:

$$ u = \frac{g_1}{s_{g_1}} \sim u_z $$

where:

$$ s_{g_1} \approx \sqrt{\frac{6}{n}} $$

(for large samples, only),
\[ g_1 = \frac{k_3}{k_2^{3/2}} \]

\( k_3 \) is the central moment of third rank, 
\( k_2 \) is the central moment of second rank, 
\( n \) is the number of observations.

Rejection of the null hypothesis \( (H_0: \gamma_1 = 0) \) indicates the possibility of a partitioning of the major locus in inheritance.

Statistics used for testing the significance of the kurtosis adopts the following form:

\[ u = \frac{g_2}{s_{g_2}} \sim u_z \]

where:

\[ s_{g_2} \text{ for large samples is } \sqrt{\frac{24}{n}} \text{ (in approximation)} \]

\[ g_2 = \frac{k_4}{k_2^2} \]

\( k_2, k_4 \) are central moments of the second and fourth ranks, respectively.

Similarly to the previous case – not rejecting the null hypothesis \( (H_0: \gamma_2 = 0) \) suggests the mixed inheritance of a trait.

From the point of view of application these methods are relatively simple, since unlike some others, they do not require the inclusion of any pedigree information. However it should be remembered that deviation from normality is not equivalent to the existence of mixed inheritance. Hence these tests are of limited usefulness in the detection of the major locus (Le Roy and Elsen, 1992).

**Merat's method**

Including the individual pedigree in evaluation of skewness and kurtosis was investigated by Merat (1968). The method is based on analysis of heterogeneity of kurtosis in fullsib (or halfsib) groups with the basic assumption that if at least one of the parents is heterozygous (e.g. Aa) the kurtosis coefficient will be negative. Such a case involves the problem of the size of fullsib or halfsib families (small number of individuals per family is connected with non-normal distribution in the groups). Hence, Merat (1968) proposed a modification of the method. The modification consists of dividing all the analysed families into two groups
according to the magnitude of their variance. Then the testing hypothesis concerning the negativity of kurtosis is employed separately for each group. Le Roy and Elsen (1992) described this modification by the following statistical formula:

$$\frac{g_{zl}}{s_{g, zl}} - \frac{g_{zh}}{s_{g, zh}}$$

It should be stressed that the division into two groups according to the size of family (or halfsib groups) variances is rather arbitrary. Population specificity is not taken into account here and thus, the efficiency of statistical inference may differ depending on the population considered.

**The method of parent-offspring regression**

This method is based on the assumed linearity of parent-offspring regression when only additive action genes are presented (Gimerfarb, 1986). The deviation from linearity may be interpreted as a result of polygenic interactions (for instance dominance and epistasis) and, above all, the gene action with large effects. Very large population size is frequently associated with numerical barriers. However, a simplified procedure may be used when instead of the pairs of parents (dam or sire) – progeny, pairs of parent-fraction of their progeny (having a trait level of the parent) are included (Snedecor and Cochran, 1980). The method was applied in investigation on a genetic determination of muscular hypertrophy in the Belgian White and Blue cattle by Hanset and Michaux (1985). It was assumed that the total variation among R proportions \( p_i \) is measured by an \( \chi^2 \) with \( (R-1) \) degrees of freedom. As there is a score \( (X_i) \) for each of these proportions, a weighted regression coefficient of \( p_i \) on \( X_i \) is calculated. The difference \( \chi^2 - \chi^2 \) is an \( \chi^2 \) with \( (R-2) \) degrees of freedom for testing the deviations of the \( p_i \) from linear regression on the \( X_i \).

**Evaluation of variance homogeneity within a family**

Many authors (Bishop et al., 1988; Elsen and Le Roy, 1990; Hill and Knott, 1990) have indicated the possibilities of applying the tests of within-family variance homogeneity for detecting genes with large effects. Thus, the within-family variance heterogeneity (as an alternative of variance homogeneity) suggested the mixed inheritance. In this case the Bartlett test (1937) is the most frequently applied. It is based on the following statistics of \( \chi^2 \):

$$\chi^2 = \frac{2.303}{c} \left[ \frac{(n - 1) \log \bar{\sigma}^2 - \frac{1}{i=1} (n_i - 1) \log \sigma_i^2}{\sum_i \left( n_i - 1 \right) \log \sigma_i^2} \right] \sim \chi^2_{\frac{c}{c-1}}$$
where:

\[ c = 1 + \frac{1}{3(t-1)} \left( \sum_{i=1}^{t} \frac{1}{n_i - 1} - \frac{1}{n-t} \right) \]

\( n_i \) is the size of \( i^{th} \) family,

\( n \) is the total number of individuals,

\( t \) is a number of families,

\( s_i^2 \) is \( i^{th} \) family variance,

\( \tilde{s}^2 \) is the general variance.

However, in some situations variance homogeneity need not be equivalent to polygenic inheritance and it is necessary to pay attention to some regularities connected with the major gene segregation. The full sibs having a mean of the trait similar to the population mean usually show a relatively large variance. However small variance of fullsib groups is usually combined with an extreme trait mean, when families contain mainly alternative homozygotive individuals, for instance AA or aa. So, in the latter case, variance homogeneity test is not an adequate method for the detection of a major locus.

**Within-family mean-variance regression method**

The above imperfection may be eliminated by the application of the criterion of mean/variance of families as described by the following curvilinear regression equation (Fain, 1978):

\[ \log \sigma_i^2 = \alpha + \beta_1 \mu_i + \beta_2 \mu_i^2 + \beta_3 \mu_i^3 \]

where:

\( \log \sigma_i^2 \) is logarithm of \( i^{th} \) family variance,

\( \alpha \) is a regression constant,

\( \mu_i \) is \( i^{th} \) family mean,

\( \beta_1, \beta_2, \beta_3 \) are regression coefficients.

Verification of the hypothesis concerning mixed inheritance is based on known statistical procedures – investigation effects in curvilinear regression by Fisher-Snedecor statistics (Mayo et al., 1980). Rejecting this hypothesis is equivalent to non-significance of components in this model.

It should be stressed that this method is usually applied in human genetics. However, simultaneous investigations carried out by Le Roy and Elsen (1992) have suggested a wider applicability of this method for the detection of the animal major locus using both fullsib and halfsib families.
MAJOR GENE INDEX (MGI)

This method is based on the intuitive assumption that under polygenic inheritance, progeny deviation from the midparent average is smaller than the deviation from either parents (Karlin et al., 1979). On the basis of this assumption, the major gene index can be presented as:

\[
MGI(k) = \frac{\sum_{i=1}^{n} \left[ O_i - \frac{1}{2} | S_i + D_i |^k \right]}{\sum_{i=1}^{n} \left( | O_i - S_i |^{k/2} | O_i - D_i |^{k/2} \right)}
\]  \hspace{1cm} [1]

where:

\( n \) is the number of sire (dam) – offspring pairs,
\( O_i, S_i, D_i \) is the observation of the \( i^{th} \) offspring, sire and dam, respectively,
\( k \) – known parameter.

The choice of value \( k \) is somewhat arbitrary. However, Karlin et al. (1979) recommend evaluation of the index at three levels of \( k \) (\( k = 0.5, 1, 2 \)). Thus calculated values of the index greater than 1 would be indicative of major gene inheritance. As previously mentioned, the opposite case (\( MGI(k) < 1 \)) indicates polygenic inheritance. This criterion of magnitude of \( MGI(k) \) for different \( k \) values as well as increasing \( k \) values also suggest major gene inheritance. When three generations are known, the formula [1] may be used after some modification (Le Roy and Elsen, 1992). There are also two other criteria for identification of major locus, the so called Structure Exploratory Data Analysis (SEDA) (Karlin and Williams, 1981): offspring between parents regression (OBP) and pairwise midparental correlation coefficient (MPCC). The general principle of using these criteria for the detection of major genes is similar to the MGI method. The application of the MGI method to detecting major genes usually requires adjustment of the phenotypes (parents and progeny) for such influences as sex, age or productive season. The classical adjustment procedure based on linear or nonlinear regression (Zuk et al., 1980) or proportional adjustment (Miller et al., 1966) is not sufficiently accurate. Hence Famula (1986) suggested a construction of the major gene index for quantitative traits using Best Linear Unbiased Prediction (BLUP) based on the mixed model:

\[
y = Xb + Zg + Zg_1 + e, \hspace{1cm} [2]
\]

where:

\( y \) is the \( m \times 1 \) vector of observations,
\( b \) is the \( p \times 1 \) vector of unknown fixed parameters,
$g_1$ is the $q \times 1$ vector of unknown random additive genetic effects,
$g_2$ is the $q \times 1$ vector of unknown random dominance genetic effects,
$e$ is the $m \times 1$ vector of random effects of errors,
$X, Z$ are the $m \times p$ and $m \times q$ known incidence matrices, respectively.

Assumptions of this model:

$$E \begin{bmatrix} y \\ g_1 \\ g_2 \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ O \\ O \\ O \end{bmatrix}$$

and

$$\text{Var} \begin{bmatrix} y \\ g_1 \\ g_2 \\ e \end{bmatrix} = \begin{bmatrix} V & ZA\sigma_a^2 & ZD\sigma_d^2 & R \\ AZ^2 & A\sigma_a^2 & O & O \\ DZ^2 & O & D\sigma_d^2 & O \\ R & O & O & R \end{bmatrix}$$

where:

$$V = ZAZ'\sigma_a^2 + ZDZ'\sigma_d^2 + R$$

$$R = I_m\sigma_e^2$$

Moreover:

$A$ is the $m \times m$ additive relationship matrix for $m$ individuals (parents and progeny),
$D$ is the $m \times m$ dominance relationship matrix for $m$ individuals (parents and progeny),
$I_m$ is identity matrix,
$\sigma_a^2, \sigma_d^2, \sigma_e^2$ are the component of additive, dominance and residual variance, respectively.

More details of the linear model and prediction with the BLUP method as well as rules of construction of these relationship matrices are given by Kennedy (1989).

Best linear unbiased predictes (the so-called animal model) of $g_1$ and $g_2$ are computed from the following mixed model equations (Henderson, 1985):

$$\begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + A^{-1} & Z'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z & Z'R^{-1}Z + D^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$
where:

\[
\lambda = \frac{\sigma^2_e}{\sigma^2_a}
\]

and

\[
\gamma = \frac{\sigma^2_e}{\sigma^2_d}
\]

The predicted additive genetic values for each individual (dam, sire and progeny) are substituted into the MGI formulae [1]. Simulation studies (Famula, 1986) indicated potentially useful applications of the index in some situations, such as, when there is a complete dominance inheritance and a low frequency of major genes (in the absence of multimodality of the phenotypic distribution). Moreover, the index is more sensitive to detection of single genes of large effects compared to phenotypic values when the predicted genetic merit is included.

The index based on mixed model methodology may be particularly useful in the following two cases: for a sex-limited trait (sire not being measured) or for the progeny-limited trait (a trait measured after slaughter).

Unfortunately, MGI is not a statistical test because control of error at a significant level is not possible. On the other hand, the method may be largely used as preliminary indicator of a major gene segregation.

SEGMENATION ANALYSIS (SA)

This method was introduced by Elston and Steward (1971) who used it to test the agreement of probabilities of single locus transmissions to Mendelian expectations. Next, Morton and MacLean (1974) proposed a new concept for segregation analysis which was based on including major locus effects as well as polygenic components and random environmental effects. Descriptions and some practical applications of these methods were discussed by Hill and Knott (1990) and Lalouel et al. (1983). Recently, the approaches both of Elston and Steward (1971), and of Morton and MacLean (1974) were modified for animal genetic problems.

"Segregation analysis" is a group of methods for major locus identification based on pedigree information, mainly of halfsib groups (Le Roy et al., 1989; Knott et al., 1990, 1992a, 1992b), fullsib groups (Le Roy et al., 1990) and all available pedigree information from the so called complex segregation analysis (Elston and Rao, 1978; Elston, 1980, 1992). Recently, multivariate segregation analysis which allows explicit tests of major gene pleiotropy hypotheses have been developed (e.g. Blangero and Konigsberg, 1991). The general principle of detecting a major gene by segregation analysis is to compare different types of models (polygenic and mixed inheritance) which include both genetic and environmental effects. Next, all pedigree information is summarized in a likelihood function which depends on different parameters (mean and residual
variance within genotype groups, heritability, genotype frequencies etc.) denoting the probability of observation, given a particular transmission hypothesis. The ratio of those likelihood functions supplies an answer to the final question concerning mixed inheritance. In this paper, the method is presented on the basis of a typical hierarchical experimental design for many domestic species, mainly poultry and pigs (Le Roy et al., 1990). It is assumed that each of n sires mated to m dams (additional, sires and dams are unrelated, and environmental effects are not significant) and each family contains l offspring. Two different transmission models have been compared. For polygenic transmission the linear model is as follows:

\[ y_{ijk} = \mu + u_i + v_{ij} + e_{ijk}, \]

where:

- \( y_{ijk} \) is the observation of \( k \)th individual,
- \( \mu \) is the overall mean of the trait,
- \( u_i \) is the random effect of \( i \)th sire, a variable distributed as a normal with mean 0 and variance \( \sigma_u^2 \),
- \( v_{ij} \) is the random effect of the \( j \)th dam mated to the \( i \)th sire; a variable distributed as a normal with mean 0 and variance \( \sigma_v^2 \),
- \( e_{ijk} \) is the random error effect normal distribution with mean 0 and variance \( \sigma_e^2 \).

Thus, this model obviously depends on the following four parameters: the general mean (\( \mu \)) and respective variance components (\( \sigma_u^2, \sigma_v^2, \sigma_e^2 \)). This model corresponds to the null hypothesis (\( H_0 \)).

In the alternative model, a major locus effect is included into the polygenic model:

\[ y_{ijk} = \mu_r + u_i + v_{ij} + e_{ijk}, \]

where:

- \( \mu_r \) is the mean of progeny with genotype \( r \),
- \( y_{ijk}, u_i, v_{ij}, e_{ijk} \) as above.

It is assumed that sex two unlinked alleles (e.g. A and a) are segregated thus three
genotypes can be obtained: AA, Aa, aa (hence \( r = 1, 2, 3 \)). Moreover, sires have
the genotypes s (s = 1, 2, 3) with probability \( p_s \). This model depends on eight
parameters: \( \mu_1, \mu_2, \mu_3, \sigma_u^2, \sigma_v^2, \sigma_e^2, p_1, \) and \( p_2 \) (\( p_3 = 1 - p_1 - p_2 \)).
It should be stressed that if other fixed effects are included in these models, the
number of estimable parameters will increase. As mentioned, the test statistics is
the following: ratio null ("polygenic inheritance" \( H_0(M_0) \)) and general ("mixed
inheritance" \( H_1(M_1) \)) hypothesis:
\[
\chi^2 = -2 \ln \frac{M_0}{M_1} \sim \chi^2_{2d}
\]

where:

d is the number of parameters with fixed values under \( H_0 \) (here: \( d \) is 1, hence degrees of freedom for central \( \chi^2 \) is 2),

\[
M_0 = \prod_{i=1}^{n} \sum_{u_i} \int f(u_i) \prod_{j=1}^{m_i} \int g(v_{ij}) \prod_{k=1}^{l_{ij}} h_0(y_{ijk}/u_i, v_{ij})
\]

and

\[
M_1 = \prod_{i=1}^{n} \sum_{s_i} \sum_{t_i} \sum_{s_i} \sum_{t_i} \int f(u_i) \prod_{j=1}^{m_i} \int g(v_{ij}) \prod_{k=1}^{l_{ij}} \sum_{r_{ijk}} P(R_{ijk} = r_{ijk}/s_i, t_i)
\]

Moreover, \( S_i \) is the genotype of the \( i \)th sire and \( s_i \) its realization, \( T_{ij} \) is the genotype of the \( j \)th dam of the sire \( i \), and \( t_{ij} \) its realization, and \( R_{ijk} \) is the genotype of the \( k \)th progeny of dam \( ij \), and \( r_{ijk} \) its realization;

\( P(R_{ijk} = r_{ijk}/s_i, t_i) \) is the probability of \( r_{ijk} \) given the genotypes \( s_i \) and \( t_i \) of the sire \( i \) and the dam \( ij \); \( f \) is the distribution of the sire effect \( U_i \):

\[
f(u_i) = \frac{1}{\sqrt{2\pi\sigma_u}} \exp \left( -\frac{1}{2} \frac{u_i^2}{\sigma_u^2} \right)
\]

\( g \) is the distribution of the dam effect \( v_i \), \( h \) is the distribution of the dependent variable, given \( u_i \) and \( v_{ij} \):

\[
h_{r_{ijk}}(y_{ijk}/u_i, v_{ij}) = \frac{1}{\sqrt{2\pi\sigma_e}} \exp \left( -\frac{1}{2} \frac{(v_{ijk} - \mu_i - u_i - v_{ij})^2}{\sigma_e} \right)
\]

Segregation analysis is a more powerful statistical test compared with the other methods mentioned above. The SA method allows for the derivation of additional results on the identified major gene, for instance, genotypic effects as well as frequency of genes and genotypes. Unfortunately, SA is a very difficult method from the numerical point of view.

The likelihood equations presented above are based on a complicated function, which gives a summation of each combination of the major genotypes for a pedigree. These calculations are impracticable even for a small number of offspring per sire (or per sire and dam — fullsib groups), since they require an extremely large computer memory. In order to simplify the calculations, several approximations to this likelihood function have been proposed. Knott et al.
(1992a), comparing three approximations in halfsib balanced experimental design, suggested using the so-called Hermite integration to replace integration in the combined model likelihood (more details of this kind of integration are given by Knott et al. (1992a, b)). Other solutions of simplified segregation analysis for one-way classification experimental design were studied by Le Roy et al. (1989). It should be stressed that the numerical complications increase if the number of observed individuals is large. In these cases, probabilities of additional relationships among individuals, overlapping of generations and large complex pedigree with loops are increasing. Thus, the calculations required are often almost impossible to complete.

Segregation analysis is also limited from the statistical point of view. Elsen and Le Roy (1989) concluded that likelihood ratio distribution has no regular asymptotic convergence properties in the case of the SA method. The likelihood ratio distribution depend on sample sizes and the structure of experimental data (nuclear families, half – or fullsib families). Hence, the main statistical test ($-2\ln M_0/M_1$) is not always true for each case. Another problem in detecting major genes by segregation analysis may be connected with the distribution of the analysed trait. As mentioned in part 2 of this paper, skewness is expected when genes with large effects are segregated. It was considered a paradox that skewness may lead to the false inference of a major gene under this method (Demenais et al., 1986). Elsen and Le Roy (1989) showed that when the skewness coefficient is larger than 0.2 the inference is quite false. Skewness can be reduced by the application of data transformations, for instance logarithmic or Box-Cox transformation (Gianola et al., 1990). However, these transformations may lead to a large loss of power (Demenais et al., 1986). Investigations carried out on both transformed and untransformed pig data by Le Roy et al. (1990) indicated similar estimates of the parameters.

REFERENCES

Bartlett M.S., 1937. Some examples of statistical methods of research in agriculture and applied biology. J. Royal Soc. Suppl. 4, 137-170


Karlin S., Williams P.T., 1981. Structure Exploratory Data Analysis (SEDA) for determining mode of inheritance of quantitative traits. II. Simulation studies on the effect ascertaining families through high valued probands. Amer. J. Hum. Genet. 33, 282-292


MAJOR ANIMAL GENES


STRESZCZENIE

Przegląd metod statystycznych stosowanych w identyfikacji genów głównych w populacjach zwierzących