Quantitative methods to improve the understanding and utilization of animal genetic resources

J.M. Ojango[1,3] N. Mpofu [1 4] K. Marshall and L. Andersson-Eklund[2]

[1]International Livestock Research Institute (ILRI), PO Box 30709, Nairobi 00100, Kenya
[2]Swedish University of Agricultural Sciences (SLU), Dept. of Animal Breeding and Genetics, PO Box 7023, SE-75007 Uppsala, Sweden (sections 1.4 and 6)
[3]Egerton University, Department of Animal Sciences, PO Box 536, Njoro, Kenya
[4]Lupane State University, Bulawayo, Zimbabwe

Quantitative methods are research techniques and methods dealing with numbers and anything that is measurable. This module reviews the most commonly used quantitative methods in the area of animal breeding and animal genetic resources (AnGR). The module primarily addresses topics of relevance to scientists in the area of AnGR in developing countries, including both faculty in universities/colleges and staff in research institutions.

The core text includes links [burgundy] and references to other parts of the AnGR Training Resources (CD), such as exercises, module texts, compendia and case studies. There are also links [blue] and references to literature and websites.

Contents

Citation

Module 4: *Ojango J.M., Mpofu, N., Marshall, K. and Andersson-Eklund L. 2011.* Quantitative methods to improve the understanding and utilisation of animal genetic resources *In:*Animal Genetics Training Resource, version 3, 2011. Ojango, J.M., Malmfors, B. and Okeyo, A.M. (Eds). International Livestock Research Institute, Nairobi, Kenya, and Swedish University of Agricultural Sciences, Uppsala, Sweden

1 Quantitative methods—Important tools for animal genetic resources (AnGR)

One of the challenges faced by livestock keepers in developing countries is the need to improve productivity per animal and per unit area of land. The most common question asked to those promoting animal genetic improvement is: 'what is the best animal?'. To determine an answer to the question of what is best, one must know the traits of importance (See Module 2, Section 4) and how performance in the traits interacts with the available environment.

A second basic question is: 'how do you breed animals so that their descendants will be better than today's animals?'. In other words, how are animal populations improved to maximize profitability over time? The purpose of animal breeding is not to genetically improve individual animals, but to improve animal populations. To improve populations, basic tools are required to identify and utilize genetic differences between animals for the traits of interest.

The vast majority of traits of interest are polygenic. The higher the number of genes that affect a given trait, the more difficult it is to observe and separate the effects of the individual genes. Phenotypes for polygenic traits are typically quantitative in their expression, and thus can be numerically measured. *Quantitative methods* are research techniques that are used to gather quantitative data—they are research methods dealing with numbers and anything that is measurable. Statistical analyses are often used to evaluate and present the results of these methods.

1.1 Statistics to separate genetic and environmental effects

The observed characteristics of animals, which make up the phenotype of the animal, are affected by both genetic and environmental factors. The genetic factors are due to a random sample of genes received from the two parental gametes, whereas the environmental factors include influences by climate, nutrition, health and management. Genetic analyses in the field of AnGR usually aim at separating genetic and environmental effects. Statistical values for means, variances and the relationship between different variances are used to develop basic analytical principles. For this, a mathematical model is needed which describes the phenotypic values as a function of genotype and environment, i.e. phenotype = f (genotype, environment). The simplest and most frequently used function f is the linear 'pattern plus residual' model. As the genotype is our main interest, we start by defining a genotypic value as G, the 'pattern' part of the phenotypic observation as P and the residual $P - G$ as an environmental effect E, explaining the discrepancy between the phenotypic and genotypic values. The simplest model to describe the above relationships is that of Falconer and Mackay (1996) presented as:

$$P = \mu + G + E$$

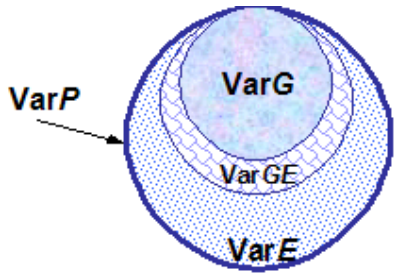
where μ = the population mean for the trait. It is the average phenotypic value of all individuals in a population. Means vary greatly with breed, management and physical environment.

The reason for adding the population mean is to emphasize that in animal breeding, genotypic values, environmental effects and all other elements of the genetic model are not absolute values but are expressed relative to the mean of the population being considered

When dealing with very different environments or when dealing with different genotypes within a given environment, it is important to include the specific combination effects between genotype and environment (G × E) (See Module 2, Section 3.4); [CS 1.39 Okeyo and Baker].

When considering breeds or populations of animals, P and G are expressed as deviations from the population mean. The variation within the population can be described and illustrated as follows:

$$\mathrm{Var}P = \mathrm{Var}G + \mathrm{Var}E + \mathrm{Var}(G \times E)$$



where Var P, VarG and VarE refer to phenotypic, genetic and environmental variance respectively and Var(G × E) is the variance due to the genotype by environment interaction. Variation refers to differences among individuals within a population. Genetic variation is the source of all genetic improvement.

Once the importance of the environmental and genetic factors for a specified trait has been established, methods of genetic improvement for that trait can be explored. Clearly, there is little or no point in attempting to improve livestock by genetic means if there is no or very small genetic variation in the trait. One must therefore determine the extent to which genetic effects influence a phenotype (Module 2 Section 3.2), and to what extent the genetic effects are due to separately acting genes, i.e. additive genetic effects, before designing breeding programmes accordingly.

With a single individual it is not possible to separate the effects of genetic and environmental factors; neither is it possible to estimate how much of its phenotypic level is due to each factor. However, with groups of livestock, estimates of the relative importance of the environmental factors, genetic factors and interaction between the two factors can be

obtained. The quantitative genetic methods reviewed in sections 1.2–1.4 provide powerful tools for analysing and handling quantitative variation in practical breeding.

1.2 Statistics to determine relationships between traits

In addition to evaluating variation within a trait, it is important to understand how two or more traits or different values may vary together, *covariation*. Knowing that any two or more genes often affect more than one trait (*pleiotropy*), sometimes in the opposite and sometimes in the same direction, the phenomenon of covariation is important. Covariation between two or more traits can also be caused by genetic linkage between loci affecting the traits. Understanding covariation helps in predicting possible effects of selection and hence in making decisions when selecting for a specific trait. The sign (positive or negative) and magnitude of the relationship between the traits to be selected must be taken into account and means sought to achieve the desirable result when it is evident that selection in one trait will cause a negative response in another related trait. The *correlation coefficient* gives a measure of the strength of the relationship between two variables (or traits). As with variance, useful correlations are phenotypic, genetic and environmental correlations. The amount of change in one variable that can be expected for a given amount of change in another variable is measured by a *regression coefficient*. This can be expressed on both the phenotypic and genotypic scales and is related to the phenotypic and genetic correlations between the two variables. The regression coefficient is useful for prediction based on other pieces of information. For example, the regression of breeding value for a trait on phenotypic value for the same trait is used to help predict an animal's breeding value based on its own performance.

1.3 Use of statistics to estimate genetic diversity from molecular data

Genetic diversity is the basis for both natural evolutionary changes and artificial selection in breeding populations. The importance of genetic diversity and measures of genetic diversity in livestock production are described in Section 3, Module 2. The section also describes molecular genetics techniques that can be used to collect data for studies on genetic diversity. Maintaining a healthy balance between adequate genetic variation in today's livestock populations in order to exploit it through selection, while aiming to achieve product uniformity and breed identity remains a constant challenge. In this module, some quantitative methods that are used to analyse molecular data to measure genetic diversity in populations (Section 7 this module) and genetic relationships between populations (Section 8 this module) are reviewed.

1.4 Statistics for deciphering effects of genes

Currently, there is much emphasis on studies of the molecular genetic background of traits in livestock. The majority of production, functional and health traits are the consequences of complex physiological systems in the animal influenced by a large number of genes and varying gene interactions. The genetic basis for such quantitative traits can neither be fully clarified nor considered in full detail, as is possible for qualitative traits, such as coat colour.

However, among all loci affecting a quantitative trait, i.e. quantitative trait loci (QTL), some contribute more and some less to the variation between individuals. Different quantitative methods have thus been developed for analysing phenotypic and molecular genetic data. Until recently, it was not possible to identify QTL, except the ones with the largest effects, the so-called major genes (Figure 1). They can be detected by segregation analysis, i.e. as deviations from the unimodal phenotypic distribution of the character. Examples of such major genes are the Culard (mh) gene causing muscular hypertrophy in cattle, the Boroola gene increasing fecundity in sheep etc.

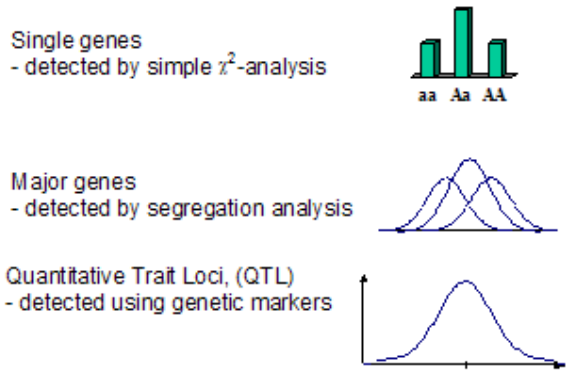


Figure 1. *The phenotypic distribution of traits influenced by different gene effects.*

As compared to studies of phenotypic distributions of traits, studies of linkage and linkage disequilibrium between genetic markers and QTL provide a more powerful and robust tool to detect QTL. Thus, the development of relatively dense linkage maps with highly informative markers (Module 2 section 3.3) has made it possible to identify and localize QTL for many economically important traits in livestock species. For the detection of QTL, it is essential to make use of efficient statistical methods (Section 6.1, this module).

2. Data collection and management

Clear objectives are required before a study is carried out. Based on these, the design, execution and analyses are planned. A key part of the research is to collect data from a number of animals. The data must be amenable to statistical analysis from which one can draw inferences or can predict future observations. There are various sources of data for use in animal breeding studies:

1. Research scientists set up experiments and collect data from experimental animals.

2. Data can be obtained from farms (field records) through livestock recording schemes (Module 3, Section 4.4).

3. Breed information data can be collected using questionnaire forms as outlined in the following section.

Scientists should have a clear understanding of the principles of statistics governing the planning of experiments and the analysis and interpretation of experimental data. It is important to design experiments properly so as to collect useful data. Costs usually prohibit the setting up and running of large experiments to collect the required data. However, often, large data sets are required to get reliable estimates of phenotypic, genetic and environmental variation.

2.1. Data sourcing through on-farm surveys of livestock breeds

In many developing countries, information on existing livestock populations in different areas is not available and livestock recording is not regularly practised. In order to understand the production systems and develop improvement programmes, it is necessary to capture existing knowledge on the livestock breeds or populations that are considered to be of most interest. Such information is generally captured through surveys. Surveys are also used to determine the status of different breeds in a country, providing key information for developing breed improvement and conservation strategies.

Designing on-farm surveys of livestock breeds

The first step is to decide what type of survey (random, purposive, convenience or representative) is to be undertaken, and the size of the population to be surveyed. Either the whole population or samples of the population can be surveyed. For a sample, the proportion of the farming community or households to be surveyed needs to be determined. This needs to be large enough to allow population values to be estimated with adequate precision; it should also cover all strata of the population related to the topic of interest. At the same time, costs of collecting data need to be realistically considered. Different sampling designs are available from simple random sampling to those using stratified and clustering techniques [Oromiya document-ILRI]. Data in surveys are usually collected using questionnaires designed to allow accurate and unambiguous answers. Key activities when carrying out a survey are illustrated in Figure 2.

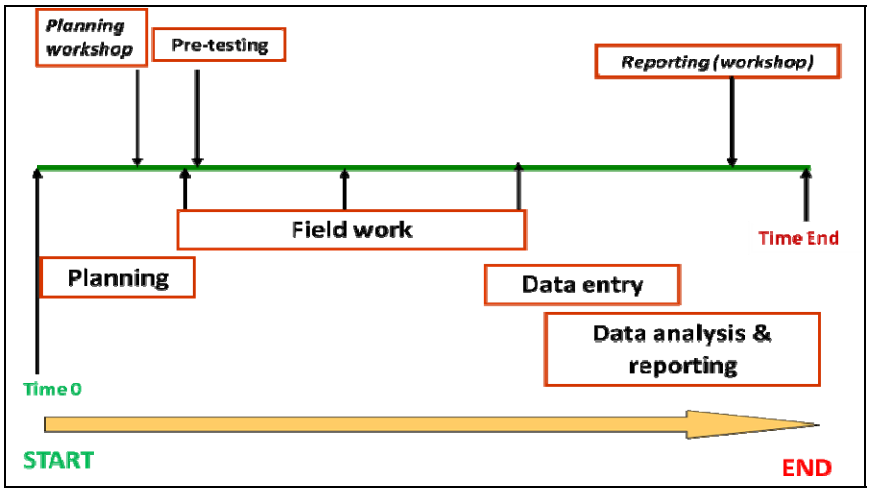


Figure 2. *Key activities when planning and implementing a breed survey*

Pre-testing of a questionnaire on a small number of farms or households is an essential and useful way of evaluating the suitability and level of detail that it is possible to obtain from the interviewees. For example, if the purpose of the survey is to estimate the population of livestock in a given area and the basic unit is a village, then one must ensure that:

- the total number of households in a village is known
- the number of such households that keep livestock is known
- the average number of livestock per livestock-keeping household is known.

These can be obtained during pre-survey visits; they give an indication of how best to achieve high accuracy and precision during the survey (see Module 2, Section 2).

Implementing on-farm surveys

In implementing on-farm surveys, the following should be considered:

- Adequate prior and mid-stream consultations with all stakeholders (farmers, local administrative officials, politicians, donors etc.)
- Timing of the survey (season and even month within seasons)
- Time for visits to farms, and where to interview respondents (in the homestead or on grazing fields)
- Who the respondents should be (household heads, children or employees).

A combination of all the above may actually be used. For example, in a society where milking is exclusively done by children and women, the best answers to the question related to how much milk an animal produces daily would be given by the family members who actually do the milking, although the household head to may respond to the entire questionnaire.

Breed descriptor charts and guidelines on animal phenotypic characteristics, such as those developed by ILRI and used for the Oromiya-ILRI Livestock Breed Survey (2001), may be available to assist enumerators and questionnaire administrators to make on-farm survey decisions. However, the occasional use of photographs to capture whole herds, while in pens, kraals or grazing, greatly helps to countercheck the accuracy and consistency of such scoring. Likewise, asking the same question to different members of the household may also help verify some discrepancies, especially where respondents seem to be giving pre-planned answers or non-plausible ones.

2.2. Data management and exploration

Raw data are entered into the computer in such a manner that the information can be found and understood long after the time of data entry, and checked for any possible errors. The data are then organized into an appropriate form for analyses. All data should be archived so that they remain available for later reference. A good data management strategy should be adapted using data management software such as Access or Oracle that has facilities for some data checking at the time of entry. Spreadsheet packages (e.g. Excel, Lotus-123) though simple and apparently flexible, should be used with caution for data management.

Data exploration

Once the necessary edits on the data have been done, it is important that one understands the data structure and the patterns displayed in the data in order to decide how best to conduct the statistical analysis [Biometrics example 1] [ICAR technical series on animal recording]. The distribution of animals by different classification (e.g. age and sex) can be determined and the mean, median and range for each factor or classification variable summarized. These statistics can then be used to group the animals into suitable subclasses to reflect the variation in the data expressed by a particular factor. Furthermore, such statistics can ensure that sufficient numbers of animals are contained within each subclass to allow reasonable inferences to be made about the influence of different levels of the factor on the trait being studied.

The number of observations per subclass usually varies for field data and some experimental data. In some cases, data that initially had an equal number of observations per subclass can end up having different numbers of observations after data editing. Data with an unequal number of observations per subclass are known as unbalanced data; there are statistical methods that have been developed to handle such data [Biometrics example 2].

In an analysis, the pattern of data is described using a model. The final model that is used to describe the data will serve as the best judge of the quality of statistical analysis. An appropriate model can only be chosen when one understands the data.

3 Statistical models for data analyses

A statistical model must, foremost, reflect the biology of the problem. A true model describes the pattern of the data perfectly but it is usually unknown. An ideal model is one that is close to a true model based on an understanding of the problem. At times, due to missing information or computational problems, an ideal model may be simplified to an operational model. This is a model that permits predictions to be made with an acceptable level of accuracy. Whenever an operational model (instead of an ideal one) is used, it is recommended that the principles for an ideal model are outlined and reasons for not using it and problems likely to arise from not using it are given. The ultimate choice of the type of model to use will depend on the traits being studied and the pattern of variation exhibited by the trait of interest.

The statistical models commonly used in animal breeding are linear models, with the set of factors being assumed to additively affect the observations. The choice of linear models has been influenced by the fact that most economically important traits studied are linear in nature (Schaeffer, 1991). More recently, non-linear models are being used to evaluate traits that exhibit categorical phenotypes (Ducrocq, 1997) and covariance functions are used in the analysis of longitudinal data (Meyer, 1998).

3.1 Components of a model

Dependent vs. independent variables: A model comprises factors/variables that influence a trait. The trait under study is termed the dependent variable, while those factors affecting it are termed independent variables. The essence of constructing a model is to determine the independent variables that affect the dependent variable, obtain information on the magnitude of each and draw inferences that can be translated into changing animal populations.

3.1.1 Characteristics of independent variables

Independent variables tend to be broadly grouped in two categories: fixed effects and random effects. Fixed effects are those estimated using information from the data only. Any conclusion drawn about the estimated mean for the trait will apply only to the study itself. They can be either discrete or continuous. Discrete factors have distinct levels, whereas continuous variables have a range of values assumed to follow a certain pattern (generally linear or quadratic). For example, it is known that calf weight at birth can be influenced by the sex of calf, the season when the dam calved, the age of the dam, the dam and the sire of calf. For the sex of calf there are two levels, i.e. male or female. Age of dam, however, can be considered as a continuous variable, say 3–12 years of age. When we fit a continuous variable, we may fit a straight line or a polynomial function of this variable. The slope of this line is known as a regression coefficient [Biometrics example 1]. Instead of treating age as a continuous variable, it is also possible to classify age of dam into different age categories

(e.g. 3, 4–6, 7–9, 10–12 years) and treat the factor as discrete with four levels [Biometrics example 2].

A covariable is a factor known to affect a performance trait which adds 'noise' to the variable of interest. When there is a significant relationship between the trait being analysed and a covariable, a proportion of the natural variation among animals is explained by this covariable. This improves the precision of comparison between mean values of primary interest [Biometrics example 2].

When a factor is considered to be random, however, results of the study can be extrapolated to a wider population from which the sample under investigation can be assumed to be drawn at random. Thus, sire, for example, is a factor that can be either fixed or random. If sires have been selected purposively for an experiment, then it is likely that we would treat the factor as fixed and calculate mean values for each sire separately. More often, though, it will be assumed that sires have been chosen at random from a wider population. In such cases the effect for sire is assumed to be random and any inferences made from the study are generalized to the wider population of which the sires are representative. To construct a model to be used in data analysis the researcher has to decide, based on the understanding of the data, whether a factor is fixed or random. As a rule of thumb, a factor is considered as random as soon as one wants to make use of prior information about the variable of interest.

3.2 Types of models used

A model comprises three parts: (1) the equation which describes the factors (effects) and their levels; (2) the specification of the distribution characteristics of random effects; and (3) assumptions, restrictions and limitations in the use of the model. There are various types of linear models. The name given depends on whether it contains only regression variables, fixed discrete effects and the number of the fixed effects in the model; whether there are any interactions between factors; or whether the model contains either only fixed or random effects or both. Thus, according to Searle (1971) and Snedecor and Cochran (1980), some of the names that one can come across are:

i. linear regression models—simple or multiple linear regression

ii. correlation models

iii. classification models—one-way, two-way, three-way classification of factors

iv. classification models with interactions

v. nested (or hierarchical) models

vi. cross-classification models

vii. random models—all factors considered random

viii. mixed models—combination of fixed and random effects

Analytical models may be for a single trait at a time (single-trait models) or for several traits at the same time (multi-trait models). When assessing several traits on the same individuals at the same time, often the interest of the researcher is to determine both phenotypic and genetic correlations between the various traits. The models used generally involve making various assumptions. An assumption often made is that residuals are normally distributed and each observation was randomly and independently obtained. However, this is not necessarily true, as if one considers a multiple trait the observations of different traits on the same animals are not independent. Also, in case of repeated measures (single trait) the observations on a same animal are not independent.

Repeated measures on an animal can cause some difficulties because adjacent observations tend to be more closely correlated relative to those further apart. Statistical procedures are generally fairly robust and slight departures from normality can be ignored. When data are clearly not distributed normally, the data should be appropriately transformed or alternative non-linear techniques can be applied.

For small data sets described by simple models (with a small number of factors), solving the equations may be quite easy. However, data sets in animal breeding can be very large and the results for the trait being evaluated can be influenced by many factors, some of which may have an uneven number of observations within each subgroup (unbalanced). For example, dairy data can include records from thousands of herds, taken over many years: some information can be missing for some herds or years. Cows within the herd can be of various genotypes and ages, cows may have been in lactation for different lengths of time etc. The statistical models required for such data sets can therefore be complicated, resulting in computational difficulties. Over time, different techniques have been developed to deal with such data, e.g. absorbing a factor to reduce the size of the system of equations to be solved, calculating the solutions of the equation system iteratively or including certain covariables or secondary factors in a preliminary step and adjusting the data for them before fitting the final model (Henderson, 1984).

Sometimes the trait of interest is measured qualitatively rather than quantitatively and observations are assigned to distinct categories or classes based on qualitative assessment of the trait. For example, cows may be diagnosed clinically as having mastitis and coded as 1 or they may be diagnosed as healthy and coded 0. Such data, when expressed as the proportion of cases occurring for different levels of a factor, often belong to a binomial, not a normal, distribution. These data do not lend themselves to direct analysis by linear models for continuous traits, although, where large amounts of data have been collected, a normal approximation can be assumed (Harville and Mee, 1984). In some cases, use of a threshold (probit) model is advisable.

4. Estimating non-genetic effects

Given the knowledge of data, a researcher will be able to develop a statistical model that describes the environmental factors likely to influence the trait of interest. For example, the environmental factors that might affect milk yield per lactation include level of herd management, level of feeding, health status of animal, age of cow at calving, season in which the cow calved etc. Some of the environmental factors will be used in the model as discrete variables, others as continuous variables. Some fixed effects influence data but are in themselves of little interest. Data are corrected for them and no explicit estimates are obtained for such factors. However, there are some effects (e.g. trends in the year or differences in sexes) whose estimates may be of interest. For these, both the fixed and random effects are estimated and predicted simultaneously using the mixed model procedures.

Parameter estimates for the fixed effects of the model are obtained most often using least squares techniques (Searle, 1971). Once the parameters have been estimated, tests can be carried out to determine whether or not the factors included in the model account for significant variation in the quantitative trait measured. The best models for evaluating fixed effects are those that take into account all the other effects in the model when estimating parameters for a given effect. In addition, estimation of linear functions and testing hypotheses related to those functions are carried out. Given the effect of season of calving, for example, one may want to test that milk yield for cows calving in the wet or cold season such as winter differs from those calving in the dry season or in the summer. The average yields for the two seasons and also differences between these yield levels can be estimated. The next step is to test the hypothesis whether seasonal differences are important for such a trait [see Biometrics example 2]. In this example, least squares analysis fitting fixed effects (discrete and continuous) is illustrated. The steps followed are: calculation of descriptive statistics, development of the model and estimation of parameters for the fixed effects.

Once the importance of environmental factors has been established, records can be corrected or adjusted for these factors before proceeding to estimate genetic effects and parameters. Procedures, such as Best Linear Unbiased Prediction (BLUP) estimate parameters for environmental factors, adjust the data for these factors and estimate genetic effects simultaneously. BLUP stands for **Best-** because it maximizes the correlation between true and predicted breeding values (or minimizes the prediction error variance); **Linear** – predictions are linear functions of observations; **Unbiased** – estimation of realized values for a random variable such as animal breeding values and of estimable functions of fixed effects are unbiased; **Prediction** – involves prediction of true breeding value [BLUP]. A variety of computer software are available at minimal or no cost to facilitate data evaluation using BLUP procedures [Computer Software].

5 Estimating genetic effects

Often, rather than estimating specific differences between treatments, one may be interested in estimating variances (phenotypic, genetic and environmental) due to the effects. For

example, in milk production the interest may not be to estimate differences between cows, but rather in estimating the variation among the cows as an estimate of the variation from a 'larger' population from which they were sampled. The cows can be considered as random effects and the data can be analysed according to a mixed effects model [Biometrics example 3].

5.1 Variance component estimation

The data described in Section 2.1 are used to estimate genetic and environmental variances needed to calculate genetic parameters (e.g. heritability) and the tests of significance for both genetic and non-genetic parameters estimated from the data.

When estimating variance components, the total variation for a trait under study is split into constituent components: genetic (additive and non-additive) and environmental. Depending on the data, different types of random effects models can be fitted. For example, dairy production data from collateral relatives (e.g. full-sibs and half-sibs) could be analysed fitting a sire model. Covariances generated by these relationships provide the information required for estimation of additive genetic variance or linear models containing both genetic and environmental effects for each animal (animal model) could be used.

The most widely used methods in variance component estimation are maximum likelihood (ML) procedures. These procedures estimate the fixed effects and variance components simultaneously. Animal breeders are increasingly confronted with data sets that have arisen from either selection experiments or from farm testing in which selection has been practised. If there is a lack of records because of selection based on some criterion that is correlated to trait(s) under analysis, the resultant estimates are likely to be biased by selection. In addition, following selection, variances of breeding values are reduced, breeding values of unrelated animals could become correlated, errors become correlated and breeding values become correlated with errors. ML statistical procedures can accommodate any structure of genetic relationship in the data, suitably weighted, do not require balanced designs and can account for selection of parents (Harville, 1977; Meyer, 1989; Falconer and Mackay, 1996).

A modified ML procedure, i.e. restricted maximum likelihood (REML) (Patterson and Thompson, 1971), has become the preferred method of analysis in animal breeding, not least for its ability to reduce selection bias. It accounts for the loss in degrees of freedom due to fixed effects in the model of analysis. In other words, it accounts for the fact that, for a given data size, more information is lost and cannot be used for estimation of variance components when one wants to estimate more levels of fixed effects.

There are several numerical procedures to find the variance components that maximize the (restricted) likelihood function, depending on whether one wants to compute only likelihood functions ('derivative-free algorithm'), first derivative also ('EM or Quasi-Newton algorithms') or first and second derivatives of these with respect to the variance components ('Newton, Fischer scoring or Average Information matrix'). Generally, impressive progress has been made in developing efficient computing algorithms for REML estimates. This,

together with increasing computing power, has enabled the analysis of quite complex statistical models in large data sets [Biometrics example 3]. There are several suites of programmes for estimation of variance components available to the scientific community free of charge, e.g. VCE (developed by Eildert Groeneveld), DMU (the Danish team in Foulum), REMLF90 (Mizstal) and WOMBAT (developed by Karen Meyer) [Web pages, Section 12, this module].

5.2 Prediction of genetic merit

There are various methods available to estimate breeding values. The quality of data will determine what method is chosen. Complete data sets will have information on performance and identity of animals. When identity and relationships are known, pedigrees can be compiled. Availability of pedigree data allows modern methods of prediction of breeding values to be used. However, to collect complete records requires that infrastructure such as identity and performance recording schemes be in place and that these schemes be well managed [CS 1.15 by Dzama]. Such schemes do not exist in most developing countries yet, and where present, financial and management constraints result in data that has a lot of missing information.

Realized values of the random variables that have been sampled from a population can be estimated if the variance–covariance structure of the population is known. The estimation of realized values of a random variable is called prediction. There are various types of predictors—best predictor (BP), best linear predictor (BLP, e.g. selection index) and BLUP (Henderson, 1984). The differences between BP, BLP and BLUP are subtle yet statistically important [van der Werf in ICAR Tech. Series No. 3].

BLUP is the most commonly used predictor to evaluate the genetic merit of livestock and in selection decisions. Several programmes that can be used for prediction of BLUP breeding values are available to the scientific community free of charge, e.g. [PEST] and [WOMBAT] and BLUPF90 (Mizstal) (see Web pages, Section 11, this module). BLUP can accommodate non-random mating and reduce bias to selection provided that the data on which selection was practised is included in the analysis. In BLUP analysis, one equation for each level of each fixed or random factor is required so that effects can be estimated simultaneously (Henderson, 1975). If there are sufficient connections between herds, as is usually the case with the use of artificial insemination, selection on BLUP values can be done on a breed (rather than herd) basis [Computer exercises: BLUP].

The various sources of information that can be used to calculate BLUP breeding values are parent and progeny, both of which are based on the pedigree and the performance of the animal.

5.2.1 Models for calculating BLUP breeding values

The ***animal model*** is now the standard method for calculating breeding values. In an animal model, the performance of an individual animal and all known pedigree relationships are used

to estimate its breeding value. The model is characterized by the fitting of a random component for the breeding value of each animal (Mrode, 2005). Use of an animal model results in a set of simultaneous equations with an order equal to the number of animals included in the analysis (with performance of their descendants), plus an additional equation for each fixed effect (Hill and Meyer, 1988). The animal model accounts for all the genetic relationships among the individuals whose breeding values are to be estimated and can account for repeated records, multiple traits, non-additive genetic effects, litter effects and a number of environmental effects, both fixed and random (Henderson, 1988). The implementation of animal models improves the correlation between proofs and true genetic values because all information is considered (Jansen, 1990; Banos et al., 1991) [Computer exercises: BLUP].

Due to computing constraints and data limitations or peculiarities, approximations or other models simpler than the animal model have been used. These include:

Sire models, where records are grouped according to the sire's identity. When using a sire model, the dams are not represented, that is they are implicitly assumed to be non-related, non-inbred and unselected. Sons of sires are accounted for in the relationship matrix between sires. Use of sire models thus leads to a downward bias in parameter estimates as only half-sib relationships are acknowledged (Henderson, 1986; Meyer, 1987).

Sire maternal grandsire models, where in addition to effects in a sire model, the effect of the dam of an animal is considered through its maternal grand sire. Here the maternal grand dams are assumed unrelated, non-inbred and unselected.

5.3 Longitudinal data analysis

Some measured traits, such as weights or milk production, are repeated over the life of the animal. It is often not adequate to consider that two such observations obtained at different ages or stages of lactation are phenotypic expressions of the same (genetic) trait. In many cases, one wants to take into account the fact that two consecutive observations are more similar than two observations far apart in time. Furthermore, the interval between measurements on the same animal may greatly vary. Therefore 'traditional discrete' multivariate models are not efficient. Such traits are called longitudinal data.

5.3.1 Random regression models

Random regression models (RRM) can be used to analyse longitudinal data. These models provide a means to estimate genetic parameters for all ages without correcting the observations to certain landmark ages (Lewis and Brotherstone, 2002; Nobre et al., 2003). The models use fixed regression coefficients to account for overall and within fixed class trends while fitting the random regression coefficients for each individual to allow for individual variations in the trajectory. For example, the genetic component of the model will be described as a polynomial function (linear, quadratic or higher order) of time. The usual assumptions (multivariate normality using a relationship matrix) are extended to all (random)

coefficients of this function. This modelling defines a particular genetic covariance between any two points in time. This continuous function that represents the variance and covariance of traits measured at different times is called covariance function (CF) (Meyer, 1998; van der Werf et al., 1998; Schaeffer, 2004). CFs are an infinite dimensional equivalent of a covariance matrix for a given number of records taken at different ages (Meyer and Hill, 1997; Huisman et al., 2002). For RRMs, the covariance function coefficients can be estimated directly by restricted maximum likelihood (REML) (Meyer and Hill, 1997; Albuquerque and Meyer, 2001).

5.3.2 Test day models in dairy production

Genetic evaluations for dairy cattle in many countries are obtained by analysing 305-day yields (or equivalent cumulative yield records) predicted from a few test-day yields (i.e. from longitudinal measurements). The 305-day yields predicted from monthly test-day records assumes such records within a single lactation measure the same trait for the whole duration of lactation. The error of genetic evaluation may further increase if 305-day yields are obtained by projecting partial lactations with factors that assume a constant shape of the lactation curve for all cows contrary to reality. Test-day records, however, are repeated observations measured along a trajectory (days in milk) and the mean and covariance between measures change gradually along the trajectory. Genetic evaluations based directly on test-day records can overcome the need to predict 305-day yields or project incomplete lactations.

Test-day models can facilitate a cheaper and more flexible recording scheme. The advantages of using these models as outlined by various authors (Stanton et al., 1992; Ptak and Schaeffer, 1993; Wiggans and Goddard, 1996; van Raden, 1997; Swalve, 1998) are:

- They can account for variable amounts of information from different lactations. By having four or more test-day yields per cow per lactation, the accuracy of a cow's genetic evaluation may be better.
- They permit estimates of fixed effects to vary across herds and stages of lactation.
- The models can describe biology and define management groups more precisely and can account for differences in the shape of the lactation curve.
- They adjust for differing effects of sampling date. The models can account for short-term seasonal effects associated with actual time of production.
- No assumptions about the 'normal' length of a single lactation have to be made.

Test-day models therefore offer an opportunity to improve the genetic evaluation of dairy cattle in tropical production situations where infrastructure to support sophisticated or detailed recording systems is limited, often resulting in data sizes too small to allow for

accurate genetic evaluation of bulls since production conditions are constrained by environment and resources (Swalve, 1998). Random regression analytical techniques are now the norm for evaluating test day yields.

5.4 Estimation of genotype by environment interactions

Tropical countries seeking to improve production levels have often imported exotic germplasm and then carried out selection in the imported population and their progeny under local conditions. This strategy is effective if production and marketing environments and selection objectives are similar for both the original and the recipient countries or production systems. However, unfavourable G × E interaction would reduce potential benefits from a strategy based entirely on continuous importation of superior germplasm from elsewhere [CS 1.16 by Mpofu]. G × E interactions are of two forms: firstly, correlations for the same trait in two environments may be significantly less than one, implying that the genetic basis for the trait differs between environments (Falconer and Mackay, 1996). The ranking of additive genetic values and hence optimal choices of selected animals may not be the same in alternative environments (Stanton et al., 1992; Calus, 2006). The second form of G × E interaction occurs when the scale of differences among breeding values for a specific trait is unequal between environments, termed 'pseudo' G × E interaction (Dickerson, 1962). In this case, the correlation between environments for true genetic value is one and the animal's ranking is the same in all environments. However, additive genetic values are lower in the more restrictive environment resulting in less response to selection [CS 1.39 Okeyo and Baker].

Cattle genotypes in diverse environments

Methods of estimating G × E are presented by Mathur and Horst (1994), Chagunda (2000), Calus (2006) and Strandberg (2006). The methods include:

- *Orthogonal comparison of subclasses*
  This method is normally used in factorial experiments. An example is when there are two genotypes raised in two environments. The interaction effect may be estimated as the difference between the sums of diagonal subclasses. The interaction is tested for significance using an F-test.

- *Factorial analysis of variance*
  For this method a linear model, with an environmental factor, a genetic factor and

interaction effect between the two factors, is fitted with genetic and interaction effects as random effects.

- *Intraclass genetic correlations*
  This procedure is based on the estimation of genetic correlations between traits measured in two environments. The requirement is that the animals in the two environments should be genetically related (Ojango and Pollott, 2002).

- *Estimation through selection in two environments*
  G × E can also be determined indirectly from direct and correlated response to selection (Falconer and Mackay, 1996). This procedure considers the problem of carry-over of improvement from one environment to the other. Selection in environment Y is based on selection in environment X. The correlated response is compared to the direct response possible through selection in environment Y. The ratio of correlated response and direct response is computed and used to calculate G × E. This method, although likely to give a reliable measure of G × E, can only be applied after selection has been practised.

- *Using reaction norm models*
  Estimating G × E in breeding value estimation can be done with a reaction norm model when the production environment can be described as a continuous variable. A norm of reaction describes the pattern of phenotypic expression of a single genotype across a range of environments. For every genotype, phenotypic trait and environmental variable a different norm of reaction can exist. Studies of heritability carried out in a single environment cannot accurately estimate the ***Norm of Reaction***, and often may not predict phenotypic response in a different environment. The reaction norm model, analysed using random regressions, has the advantage that no arbitrary grouping of environments is required and it can be extended to handle multiple environmental scales and multiple traits (Calus, 2006; Strandberg, 2006).

5.5 Estimating heterosis effects

Cross breeding is a popular method of genetic improvement of livestock, especially in developing countries where previously such practices have been mostly inappropriately designed or executed [CS 1.34 Panandam and Raymond]. The basis of the effects and benefits derived from systematic cross breeding can broadly be classified into additive and non-additive. The additive component is that which is due to the averaging of the additive merit in the parental breeds with simple weighting according to level of gene representation of each parental breed in the crossbred genotype (Swan and Kinghorn, 1992). Heterosis is the non-additive effect of cross breeding. It is the amount by which merit in crossbreds deviates from the additive component. Heterosis is usually attributed to genetic interactions within loci (dominance) and between loci (epistasis). Individual heterosis is the deviation in performance in an individual relative to the average of the parental breeds, whereas maternal heterosis refers to heterosis attributed to using crossbred instead of purebred dams and occurs due to the dam itself possessing heterosis [CS KDPG].

The performance of crosses can be predicted using estimates of genetic parameters from cross breeding experiments. Models for estimating cross breeding parameters based on a two-locus factorial model of gene effects were developed first by Dickerson (1973) and later by Küttner and Nitter (1997). A case study by Kahi [CS 1.5 by Kahi] illustrates an example of data analysis for estimating cross breeding parameters for milk production traits under the humid coastal regions of East Africa, while another by Aboagye [CS 1.9 by Aboagye] gives such parameters for milk production, reproductive, growth and carcass traits in cattle under the humid West African tropical conditions. Software such as CBE (cross breeding effects) are also available that be used to estimate cross breeding effects from a larger variety of data structures or experimental designs.

5.6 Analysis of ordered categorical traits

Traits such as calving ease or litter size are expressed and recorded in categories. For example, in the case of calving ease, births may be assigned to one of several distinct classes such as difficult, assisted and easy calving. Usually, these categories are ordered along a gradient. In the case of calving ease, for example, the responses are ordered along a continuum measuring the ease with which birth occurred. These traits are therefore termed ordered categorical traits. Such traits are not normally distributed and animal breeders have usually attributed the phenotypic expression of categorical traits to an underlying continuous unobservable trait which is normally distributed, referred to as the liability (Falconer and McKay, 1996). The observed categorical responses are therefore due to animals exceeding particular threshold levels of the underlying trait.

Linear and non-linear models have been applied for the genetic analysis of categorical traits with the assumption of the underlying normally distributed liability. Usually, the non-linear (threshold) models are more complex and have higher computing requirements. The advantage of the linear model is the ease of implementation as programs used for analysis of usual quantitative traits could be utilized. However Fernando et al. (1983) indicated that some of the properties of BLUP do not hold with categorical traits. In a simulation study, Meijering and Gianola (1985) demonstrate that with no fixed effects and constant or variable number of offspring per sire, an analysis of a binary trait with either a linear or non-linear model gives similar sire rankings. This was independent of the heritability of the liability or incidence of the binary trait. However, with the inclusion of fixed effects and variable number of progeny per sire, the non-linear model gave breeding values more similar to the true breeding values compared with those estimated using the linear model. The advantage of the threshold model increased as the incidence of the binary trait and its heritability decreased. Thus for traits with low heritability and low incidence, a threshold model might be the method of choice. Further information on these can be found in Mrode (2005).

6 Mapping quantitative trait loci

6.1 Strategies for QTL analyses

The aim of QTL analyses is to detect, localize and estimate effects of QTL. The principle of the analyses is to search for non-random associations between phenotypic records and chromosome segments across the genome. Within the segments, the genetic constitution of each animal is deduced from the inheritance of genetic markers. Significant differences in phenotypic expressions between animals with different genetic constitutions indicate the existence of QTL in the studied chromosome segment. For example, consider the simple case of a large half-sib family, whose sire is heterozygous for a QTL and a marker near that QTL (e.g. Q—M and q—m). Offspring inheriting the 'M' marker allele (and thus mostly the QTL allele 'Q') will have a different mean to those inheriting the 'm' marker allele (and thus mostly the QTL allele 'q').

In some cases, candidate genes for QTL are known based on information from other populations or other species. Known candidate genes can be tested directly using polymorphisms within the gene or markers closely linked to the gene.

When the aim is to detect unknown QTL, an initial scan of the entire genome has to be performed. In this case markers are genotyped at roughly even spacing across the genome. The genome scan can show the chromosome segments in which QTL are located, but the accuracy of the location is usually low. To increase the precision, and thus improve the possibilities of identifying the QTL, the chromosome segments of interest need to be further studied using other methods, i.e. fine mapping.

All phases of QTL mapping (Figure 3) involve analyses of quantitative traits that have a complex genetic background and are influenced by environmental factors. Therefore, in addition to the need for genetic marker information, powerful analyses require good phenotypic records from a large number of animals and the use of suitable quantitative statistical methods.

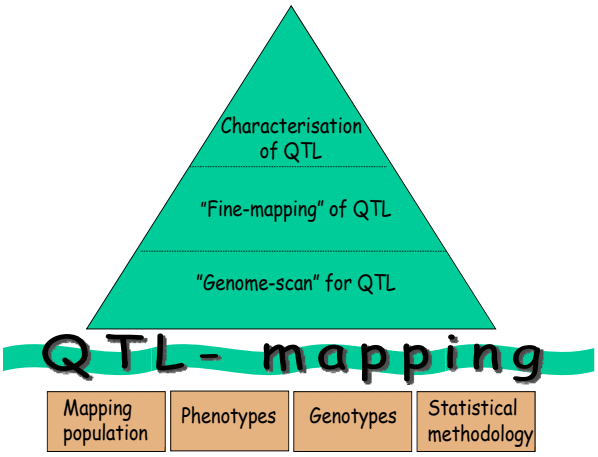


Figure 3. *The phases of QTL mapping*

A full genome scan for QTL, aimed at finding the approximate QTL location for subsequent fine mapping and possible use in marker assisted selection (MAS), includes the following steps:

i. *Choice of a mapping population*: In domestic animals we can either use experimental crosses between divergent populations (such as a breed susceptible to a disease crossed with a breed resistant to the disease) or large families within a population. Studies in designed crosses (e.g. back-cross or intercross designs) are powerful, as they help ensure that family parents (e.g. sires) are heterozygous for the QTL. However, such experiments are expensive for large animals and they do not give any direct answers in relation to the segregation of QTL within the commercial population of interest. The use of families within a population (e.g. a half-sib design) has the advantage that detected QTLs will segregate within the commercial population, but the disadvantage that all sires may not be heterozygous for the QTL.

ii. *Collection of phenotype data*: To ensure the analysis has sufficient power to detect the QTL(s) of interest, phenotypes are required on a large number of animals. They can either be the same animals that are genotyped or offspring of the genotyped individuals (progeny testing).

iii. *Genotyping*: Genetic maps, based on DNA markers, are available for many species (see http://www.ncbi.nlm.nih.gov/mapview/). Amongst others, the DNA markers include microsatellites (which are short tandem repeats) and single nucleotide polymorphisms (SNPs; point mutations in the genome). For the genome scan a subset of informative, evenly spaced markers covering the entire genome is selected for the population of interest. The maximum distance between the markers depends on the size of the population and the size of the QTL effects to be detected.

iv. *Setting up a genetic model for QTL*: Depending on data available, an operational model with one or several QTL (with additive, dominance, epistatic or substitution effects) and remaining genetic and environmental effects is used.

v. *Drawing statistical inference from data*: The statistical testing for QTL can be performed at marker loci (single marker analysis) or at marker loci and in intervals between markers (interval mapping). In practice, interval mapping is typically used, as in single marker analysis the recombination frequency between the marker and the QTL and the size of the QTL effect are confounded. Different methodologies to test for QTLs include regression, ML and variance component models, amongst others. Due to multiple testing across the genome, permutation testing is typically used to set significance thresholds.

Genome scans usually locate putative QTL to a wide chromosomal region (e.g. 30 to 60 cM). For this reason a genome scan may be followed by a fine-mapping experiment which aims to reduce the confidence region around the QTL to less than a few cMs. Fine mapping typically involves: a) typing more closely spaced markers within the region of interest; b) increasing the experimental population size; and c) use of alternate mapping methodologies (such as approaches based on linkage disequilibrium (LD)). In turn, fine mapping may be followed by experiments aimed at detecting the actual gene of interest and the causative mutation(s) within the gene (see Grisart et al., 2002 as a case study).

Useful references on QTL mapping methodologies include the Armidale Animal Breeding Summer Course notes (2003) and van der Werf et al. (2007).

6.2 Genome-wide association analysis

QTL mapping has been successful because of the possibilities to carry out sufficiently large experiments to give a reasonable statistical power for QTL detection. To date, thousands of QTLs have been reported (http://www.animalgenome.org/QTLdb/). However, the identification of the underlying causative mutations remain challenging. Genome-wide association analysis (GWAA) provides a new approach for high resolution genetic analysis, thanks to the development of large panels of SNPs and the development of cost-effective methods for large-scale SNP genotyping and analysis. The number of SNPs required for GWAA depends on the patterns of linkage disequilibrium in the population. Although most domestic animals are not highly inbred, their population structure makes them appropriate for GWAA because they resemble to some extent recombinant inbred lines. Breeds have been developed from large populations by dividing them into many smaller often closed populations on the basis of specific traits. This has led to a reduced genetic diversity within breeds and large haplotype blocks. GWAA surveys most of the genome for causal genetic variants. Because no assumptions are made about the genomic location of the causal variants, the approach could exploit the strength of associations between individual SNPs and phenotypes without having to guess the identity of the causal genes. GWAA therefore presents an unbiased yet fairly comprehensive approach that can be attempted even in the absence of convincing evidence regarding the function of a location of the causal gene. One fundamentally different approach 'admixture mapping' could also gain prominence in unravelling the genetic basis of complex traits in domestic animals.

6.3 Genomic selection

Recently there has been interest in an approach termed 'genomic selection' (GS) as an alternative to the above to identify chromosomal regions of interest for subsequent use in selection decisions. Under GS tens of thousands of SNP markers, closely spaced across the genome such that most or all QTL are in linkage disequilibrium with one marker, are tested for non-random associations with phenotypic records. This is usually performed in a large population—often >1000 individuals representing a large number of families. Useful references for GS methodologies include the Armidale Animal Breeding Summer Course notes (2008), and Goddard and Hayes (2007).

6.4 Why map QTL?

The detection and localization of QTL is valuable for several reasons. Firstly, we still know very little about the genetic background of quantitative traits such as growth, muscular development, milk yield, disease resistance etc. Mapping of QTL gives us better insight into the action and interaction of individual genes, which will give us opportunities to refine the genetic models used to describe the variation in quantitative traits. Secondly, associations between genetic markers and QTL can be utilized to improve the efficiency of selection

schemes, although this has found limited utility in practice (see Module 3, Section 4.7; Marshall et al., 2009 for a discussion of marker based selection in relation to developing countries). In the case of GS a prediction equation is estimated so that selection of candidates in subsequent generations can be based on genotype information only. Thirdly, mapping of QTL will eventually allow us to identify some of the genes and to study the molecular biology underlying the traits. This knowledge may in the near future be used for genetic modification of genes that are important in breeding programmes, for development of efficient vaccines etc.

7 Measuring genetic diversity from molecular data

7.1 Determining genetic structure and genetic variability between and within breeds

To understand the influence of selection, mating systems and other breeding interventions in population genetics, it is important to describe and quantify the amount of genetic variation in a population and the pattern of genetic variation among populations. Genetic variation may be measured at various levels, e.g. allelic variation at structural loci (see Module 2, Section 3). Genetic variation within breeds decreases as a result of selection for economically important traits yet genetic variation between and within breeds is important as raw material for genetic improvement. Populations showing a great deal of variation will be able to adapt to changing circumstances whereas populations with less genetic variability will be less adaptable to sudden environmental changes.

7.1.1 Allele frequency determination and allelic variability

The frequencies of an allele at loci are calculated manually by direct counting. The mean number of alleles (MNA) observed over a range of loci for different populations is considered to be a reasonable indicator of genetic variation. This holds true provided that the populations are at mutational-drift equilibrium and that the sample size is almost the same for each population. Breeds with a low MNA have low genetic variation due to genetic isolation, historical population bottlenecks or founder effects. A high MNA implies great allelic diversity which could have been influenced by cross breeding or admixture. Bar charts can be created for individual breeds to show variability in allelic distributions at loci. Given that sample sizes are never the same for each population analysed, other indicators of allele variability include the effective number of alleles (ENA) and allelic richness (Ar). ENA denotes the number of equally frequent alleles it would take to achieve a given level of gene diversity. It allows one to compare populations where the number and distribution of alleles differ drastically. Ar, however, is a measure of the number of alleles per locus but allows comparisons to be made between samples of different sizes by using the rarefaction technique or a Bayesian simulation approach to standardize populations to a uniform sample size.

7.1.2 Variation in gene frequencies

The variation in gene frequencies at each locus can be used to determine genetic variability between breeds. Chi square analysis is used to test differences among loci and breeds.

7.1.3 Variation in genotype frequencies

Variability between breeds can be measured using the observed genotypes at each locus and between pairs of breeds. The assumption of independent distribution of genotypes over all breeds can be tested by contingency Chi square analysis. Comparisons between pairs of breeds are performed.

7.1.4 Testing for Hardy-Weinberg equilibrium

Most deductions about populations and quantitative genetics depend on the relationship between gene frequencies and genotype frequencies. A population is said to be in Hardy-Weinberg equilibrium (HWE) when gene and genotype frequencies remain constant from generation to generation. There are factors which can cause changes in these frequencies (e.g. selection, migration and mutation) resulting in non-random union of gametes. Deviation from HWE in a population indicates possible inbreeding, population stratification and sometimes problems with the genotyping. In populations where individuals may be affected by particular ailments or may be under different selective pressures, these deviations can also provide evidence for association. The data required to perform HWE tests are gene and genotype frequencies and the size of sample population at each locus.

The deviation from HWE can be tested using any one of the following three methods:

a. The Chi square statistic for asymptotic tests has been used to evaluate the overall discordance of genotype frequencies at each locus or population combination (Hammond et al., 1994; Deka et al., 1995). The test is performed for every breed at each locus.
b. The likelihood ratio test criterion (G statistic) has also been used to contrast observed and expected genotype frequencies (Hammond et al., 1994; Deka et al., 1995).
c. The third method uses an exact test of HWE (conditional exact test which is analogous to Fisher's exact test for contingency tables). In addition, for loci or population combinations with five or more alleles, a Markov chain algorithm is used to obtain an unbiased estimate of the exact probability of being wrong in rejecting HWE. This method should be preferred for small sample sizes and multi-allelic loci since the Chi square test is not valid in such cases.
d. Recently, there has been great interest in testing for HWE in GWAA in which departures from HWE may indicate problems with quality control for the SNP in question. Therefore, a fourth recently derived method is based on Bayesian simulations and performs an exact test on the basis of the comparison between weighted likelihoods under the null and alternative hypotheses. The ratio of these two functions gives the Bayes Factor (BF). A distribution of the BF under the null hypothesis defines a natural order in the sample space. The discreteness of the sample space causes no complications for the Bayesian approach because all inferences are conditional on the configuration of the observed counts which negates the need to consider hypothetical data realizations. Therefore the test is exact and unconditional

and does not depend on asymptotic results. In addition, the test is desirable in terms of decision theory, as it minimizes a linear combination of Type I and type II errors.

With the exception of the Bayesian approach, GENEPOP, FSTAT, ARLEQUIN and the R-programming language can be used to test for HWE.

7.1.5 Estimating average heterozygosity

Heterozygosity is a measure of genetic variation within a population. High heterozygosity values for a breed may be due to long-term natural selection for adaptation, to the mixed nature of the breeds or to historic mixing of strains of different populations. A low level of heterozygosity may be due to isolation with the subsequent loss of unexploited genetic potential. Locus heterozygosity is related to the polymorphic nature of each locus. A high level of average heterozygosity at a locus could be expected to correlate with high levels of genetic variation at loci with critical importance for adaptive response to environmental changes (Kotzé and Muller, 1994).

The observed heterozygosity is defined as the percentage of loci heterozygous per individual or the number of individuals heterozygous per locus. Average heterozygosity at each locus and for each breed can be estimated from allele frequencies at each locus. Individual breed average heterozygosity is estimated by summing heterozygosities at each locus and averaging these values over all loci. Locus heterozygosity is estimated by summing the heterozygosity at all loci for each breed and averaging this quantity over all breeds. The expected heterozygosity (also called gene diversity) is calculated from individual allele frequencies (Nei, 1987). The FSTAT (Goudet, 1995), GENETIX (Belkhir et al., 1996-2004), R-package, Microsatellite Analyzer (Dieringer and Schlštterer, 2003) and MSTollkit (Park, 2001) computer programs can be used to estimate both observed and expected heterozygosity per locus and population and across all populations analysed.

7.1.6 Estimating levels of inbreeding

Molecular data can also be used to estimate inbreeding values even though there are factors other than descent for two markers to be similar. Observed and expected heterozygotes at different loci can be used to estimate the extent of inbreeding. The locus inbreeding coefficients are averaged to estimate average inbreeding coefficients for each population. Inbreeding coefficients should only be estimated for breeds which show significant deviation from the HWE. A large value reflects the existence of a small number of heterozygote genotypes and an excess of homozygote genotypes. A small value indicates the occurrence of heterozygote genotypes at a higher proportion than the homozygote genotypes.

7.1.7 Genetic differentiation

Population differentiation can be assessed by determining whether allelic composition is independent of population assignment (Raymond and Rousset, 1995a). The statistical test is based on analysis of contingency tables using a Markov Chain procedure to derive an unbiased estimate of the exact probability of being wrong in rejecting the null hypothesis, i.e.

allelic composition is independent of population assignment (no differentiation). The test is performed for pair-wise inter-population comparisons on contingency tables containing data from each of the microsatellite loci studied. The FSTAT, GENETIX and POPULATIONS statistical program's can be used to perform the computations.

7.1.8 Analysis of gene flow, genetic admixture and structure

a. *Use of diagnostic allele* Diagnostic alleles are alleles that are unique to certain breeds, e.g. alleles unique to indicine breeds or taurine breeds. They are used to determine the purity of breeds, the introgression by one breed type into a population and to determine the genetic composition of breeds. The frequencies of the diagnostic alleles or groups of alleles at a particular locus are averaged to give an estimate of the frequency of the diagnostic alleles in each population.

b. *Estimation of genetic admixture proportions from allele frequencies* Genetic admixture proportions can be estimated directly using a method developed by Chakraborty (1985) which uses the concept of gene identity coefficient—the probability that two genes chosen at random from one or more populations are identical in state. The underlying rationale to this method is that genetic similarity between populations can be expressed as a simple linear function of admixture proportions. This method requires that parental populations represent the original populations that produced the dihibrid populations of interest. An example would be an Asian breed (or group of Asian breeds) representing an indicine population and a group of African breeds representing a taurine population.

   A computer program called ADMIX (Chakraborty, 1985) uses a vector-matrix approach to produce weighted least squares solutions for each individual admixture proportion with associated standard errors. It also produces correlation coefficients for the weighted least squares solutions that give an indication of the validity of the underlying admixture model (i.e. do present-day Asian zebu and the African breeds serve as adequate surrogates for the original parental populations).

   Another program called GENECLASS 2.0 (Piry et al., 2004) employs multilocus genotypes to select or exclude populations as origins of individuals (assignment and detection of migrants). Both of these tests compute likelihoods using Bayesian simulations, allele frequency data or genetic distances between individuals to assign individuals to their populations of origin or detect recent immigrants.

c. *Evaluating the genetic structure of populations*
   The inherent genetic structure of populations can be assessed directly using a method developed by Pritchard et al. (2000) and implemented in the program STRUCTURE. The program implements a model-based clustering method to infer population structure, assign individuals to populations and identify migrants and admixed individuals using multilocus genotype data independent of prior population information. The approach implemented in STRUCTURE assumes a model in which there are *K* populations (where *K* may be unknown), each of which is characterized by

a set of allele frequencies at each locus. Individuals in the sample are assigned probabilistically to populations or jointly to two or more populations if their genotypes indicate them to be admixed.

7.1.9 Tests for linkage disequilibrium

Linkage disequilibrium (LDE) is the non-random association between different loci which may arise from: (i) admixture of populations with different gene frequencies; (ii) chance in small populations (e.g. endangered breeds); (iii) selection favouring one combination of alleles over another; or (iv) the close association between markers in the same linkage group (Falconer and Mackay, 1996). A test can be carried out to check for the existence of the association between markers studied. The null hypothesis for the LDE test is that all the genotypes at one locus are independent of those at another locus. The GENEPOP program (Raymond and Rousset, 1995b) and FSTAT (Goudet, 1995) can be used to test for LDE. The program prepares contingency tables for all pairs of loci in each population and in a pooled sample of all populations. Then a probability test (or Fisher exact test) for each table using the Markov chain method to obtain P-values is performed.

7.1.10 Distribution of genetic diversity (population differentiation)

When a population is divided into subpopulations, there is less heterozygosity than there would be if the population was undivided. Founder effects acting on different subpopulations generally lead to subpopulations with allele frequencies that are different from the larger population. Since allele frequency in each generation represents a sample of the previous generation's allele frequency, there will be greater sampling error in these small groups than there would be in a larger undifferentiated population. Hence, genetic drift will push these smaller demes toward different allele frequencies and allele fixation more quickly than would take place in a larger undifferentiated population. There are two commonly used approaches to quantify the distribution of genetic diversity within and between populations.

a. Wright's *F statistics*
   The decline in heterozygosity due to subdivision within a population has usually been quantified using an index known as Wright's *F* statistic, also known as the fixation index. The *F* statistic is a measure of the difference between the mean heterozygosity among subdivisions in a population, and the potential frequency of heterozygotes if all members of the population mix freely and non-assortatively (Hartl and Clark, 1997). The fixation index ranges from 0 (indicating no differentiation between the overall population and its subpopulations) to a theoretical maximum of 1. In practice, however, the observed fixation index is much less than 1 even in highly differentiated populations. Fixation indexes can be determined for differentiated hierarchical levels of a population structure, to indicate, for example, the degree of differentiation between sub-populations within a population, between populations within a group and between groups of populations. To determine the fixation index, the mean heterozygosity at each level must be determined.

b. AMOVA (Analysis of molecular variance)

The most commonly used programs for performing AMOVA are Arlequin, GDA and GenAlEx. To perform AMOVA, a distance matrix is created within any of the above programs or included within the input file. For example, Arlequin partitions the sum of squared deviations from the distance matrix into hierarchical variance components which are tested for significance using permutation tests. The AMOVA approach used in Arlequin is essentially similar to other approaches based on analyses of variance of gene frequencies, but *for certain types of data it can also take into account the number of mutations between molecular haplotypes* (**Φ**; see p 65 of manual and Excoffier et al., 1992).

- For haplotypic data, Arlequin estimates Φ using information from both the allelic content and frequency of haplotypes (Excoffier et al., 1992).
- For genotypic data, with an unknown gametic phase (as is the case for most natural populations) the AMOVA is based on F-statistics.

AMOVAs can be used to: (1) describe the partitioning of genetic variation among and within groups; and (2) test user-defined groupings of populations. AMOVA differs from a simple analysis of variance (ANOVA) in that data are arranged hierarchically and mean squares are computed for groupings at all levels of the hierarchy. This allows for hypothesis tests of between-group and within-group differences at several hierarchical levels.

8 Genetic relationships between populations

Multivariate analysis is used to describe analyses of data sets for which more than two observations or variables are obtained for each individual or unit studied. For genetic diversity studies, gene frequencies can be determined for several loci in several breeds or populations. Multiple regression and multiple correlation procedures are multivariate techniques which have had the greatest application in animal breeding research. However, these techniques are not suitable when the number of observations or variables is large. Cluster analysis and principal component analysis are two multivariate methods that have been used to analyse data generated by molecular genetics studies [CS 1.10 by Okomo]; [CS 1.11 by Gwakisa].

8.1.1 Cluster analysis

Clustering is a technique for grouping individuals into unknown groups to assess the relationship between the groups (e.g. livestock populations). With cluster analysis the number and characteristics of the groups are to be derived from the data and are not usually known before the analysis. In animal diversity studies, cluster analysis has been used to classify breeds or strains into groups on the basis of their genetic characteristics. Some initial analysis is usually recommended before clustering. Common initial analyses include scatter diagrams, profile analysis and distance measures. Scatter diagrams and profile analysis fail when the number of observations is large. For a large data set, distance measures are more appropriate.

They define some measure of closeness or similarity of two observations. In animal breeding, distance measures are called genetic distance.

a. *Genetic distance estimates*

   Genetic distances give the extent of gene differences between populations (and hence genetic relationships among them) measured by some numerical quantity and usually refer to the gene differences as measured by a function of gene frequencies. There are several measures of genetic distances. In most situations, different distance measures yield different distance matrices, in turn leading to different clusters. Examples include the standard genetic distance developed by Nei (1972), and a genetic distance measure developed by Goldstein et al. (1995). The efficiencies of the various measures of genetic distances are compared in Takezaki and Nei (1996). Several computer programs are now available for estimating genetic differences, for example, DISPAN (Ota, 1993) (see Section 12, this module).

b. *Phylogenetic analysis*

   The commonly used methods of clustering fall into two general categories: hierarchical and non-hierarchical. Hierarchical procedures are the most commonly used in animal diversity studies. When the number of variables is more than two and the data set is large, dendrograms have been used. In a dendrogram, the horizontal axis lists the observations in a particular order. The vertical axis shows the successive steps or cluster numbers.

In animal diversity studies, hierarchical procedures are called phylogenetic analysis. The genetic distance measures are used to construct the dendrograms, also called phylogenetic trees. The two most commonly used methods for constructing the trees are unweighted pair group method (UPGMA) and the neighbour-joining method (NJ) (Saitou and Nei, 1987). The operational taxonomic units (OTUs) in breeding are livestock populations or breeds. Therefore, the phylogenetic trees summarize evolutionary relationships among breeds or populations and categorize cattle populations into distinct genetic groups. The trees consist of nodes and branches. The nodes are the breeds and the branch lengths between breeds are graphical estimates of the genetic distance between the breeds and give an indication of genetic relationships between breeds. UPGMA trees give an indication of the time of separation (divergence) of breeds. The higher the branch length the longer is the separation period between breeds [CS 1.10 by Okomo]; [CS 1.11 by Gwakisa]. Bootstrapping is usually done to provide confidence statements about the groupings of the breeds as revealed by the dendrograms and hence test the validity of the clusters obtained. The bootstrap values are given in percentages and the higher the value, the higher is the confidence in the grouping. Programs such as SAS (Statistical Analysis System) and SPSS can produce dendrograms.

There are some problems with hierarchical procedures. An undesirable early combination can persist throughout the analysis and may lead to artificial results. It may then become necessary to perform the analysis several times after deleting certain suspect observations.

For large sample sizes, the printed dendrograms become too large and unwieldy to read. Another important problem is how to select the number of clusters. No standard objective procedure exists for making the selection. The distance between clusters at successive steps may serve as a guide. In addition, the underlying situation may suggest a natural number of clusters.

8.1.2 Principal components analysis

Principal components analysis (PCA) provides a method of explaining the covariance structure among a large system of measurements by generating a smaller number of artificial variates. In this manner, principal components can be used objectively to evaluate variation in measurements and to increase understanding of structural relationships as an entity rather than as a series of individual and independent relationships. In PCA, the variables are treated equally as opposed to being divided into dependent and independent variables, as is done in regression analysis. The original variables are transformed into new uncorrelated variables that are called principal components (PC). Each PC is a linear combination of the original variables. The initial variates are replaced with a smaller number of latent variates (the PC) allowing the data to be summarized more concisely with minimal loss of information. Thus, instead of analysing a large number of original variables with complex interrelationships, the investigator can analyse a smaller number of uncorrelated PCs (Morrison, 1976).

One of the measures used to determine the amount of information conveyed by each PC is its variance (usually known as eigenvalue). For this reason, the PCs are arranged in order of decreasing variance. Thus, the most informative PC is the first and the least informative is the last while a variable with zero variance does not distinguish between the members of the population. To reduce the dimensionality of a problem, only the first few PCs are analysed. The PCs not analysed convey only a small amount of information since their variances are small. The number of components selected may be determined by examining the proportion of total variance explained by each component. The cumulative proportion of total variance indicates, to the investigator, just how much information is retained by selecting a specified number of components. Ideally, we wish to obtain a small number of PCs which explain a large percentage of the total variance. Once the number of PCs is selected, the investigator should examine the coefficients defining each of them to assign an interpretation to the components. A high coefficient of a PC on a given variable is an indication of high correlation between that variable and the PC. PC scatter graphs are drawn by plotting the PC coefficients. Two- and three-dimensional scatter graphs have been used. Related breeds are clustered together.

The PCA procedures in genetic studies were described by Cavalli-Sforza et al. (1994). In animal genetic diversity studies, PCs have been used to determine relationships among populations, supplementing relationships determined using phylogenetic analyses (e.g. Okomo, 1997). PCs can be more convenient than phylogenetic trees if clusters of populations are more visible. They are also more flexible than trees since they can use a greater number of parameters. It is usually easier to compare PC maps than it is to compare trees.

9 References

Albuquerque, L.G. and Meyer, K. 2001. Estimates of covariance functions for growth from birth to 630 days of age in Nelore cattle. *Journal of Animal Science* 79:2776–2789.

Banos, G., Schaeffer, L.R. and Burnside, E.B. 1991. Genetic relationships and linear model comparisons between United States and Canadian Ayrshire and Jersey bull populations. *Journal of Dairy Science* 74:1060–1068.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. and Bonhomme, F. 1996–2004 GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France).

Calus, P.M. 2006. *Estimation of genotype × environment interaction for yield, health and fertility in dairy cattle*. Ph.D. thesis, Animal Breeding and Genetics, Wageningen University, Wageningen, The Netherlands.

Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. 1994. *The history and geography of human genes*. Princeton University Press, Princeton, New Jersey, USA.

Chagunda, M.G.G. 2000. *Genetic evaluation of the performance of Holstein Friesian cattle on large-scale dairy farms in Malawi*. PhD thesis, Georg-August-University, Göttingen, Germany.

Chakraborty, R. 1985. Gene identity in racial hybrids and estimation of admixture rates. In: Neel, J.V. and Ahuja, Y. (eds), *Genetic microdifferentiation in man and other animals*. Indian Anthropological Association, Delhi, India. pp. 171–180.

Deka R.L., Shriver, M.D., Yu, L.M. and Decroo, S. 1995. Population genetics of dinucleotide $(dC\text{-}dA)_n.(dG\text{-}dT)_n$ polymorphisms in world populations. *American Journal of Human Genetics* 56:461–474.

Dickerson, G.E. 1962. Implications of genetic–environmental interaction in animal breeding. *Animal Production* 4:47–63.

Dickerson, G.E. 1973. Inbreeding and heterosis in animals. In: *Proceedings of the animal breeding and genetics symposium in honour of Dr J.L. Lush*. American Society of Animal Science and Dairy Science Association, Champaign, Illinois, USA. pp. 54–77.

Dieringer, D. and Schlštterer, C. 2003. Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes* 3(1): 167–169.

Ducrocq, V. 1997. Survival analysis, a statistical tool for longevity data. In: *48th Annual meeting of the European Association of Animal Production*. Vienna, Austria. pp. 1–14,

Excoffier, L., Smouse, P. and Quattro, J. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.

Falconer, D.S. and Mackay, T.F.C. 1996. *Introduction to quantitative genetics*. 4th Edition. Longman Publishing Group, London, UK. 160 pp.

Fernando R. L. Billingsley, R.D. and Gianola D. 1983. Effects of method of scaling on heritability estimates and sire evaluations for frame size at weaning in angus cattle Journal of Animal Science 56:1047-1056

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222–231

Goddard, M. E., and B. J. Hayes. 2007. Genomic selection. Journal of Animal Breeding and Genetics 124:323–330

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. and Feldman, M.W. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139:463–471.

Goudet J. 1995. Fstat version 1.2: a computer program to calculate Fstatistics. *Journal of Heredity* 86(6):485–486.

Hammond, H.A., Jin, L., Zhong, Y., Caskey, C.T. and Chakraborty, R. 1994. Evaluation of 13 short tandem repeats loci for use in personal identification applications. *American Journal of Human Genetics* 55:175–189.

Hartl D.L., Clark A.G. 1997. Principles of population genetics. Sinauer Associates, Sunderland, MA

Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association* 72:320–338.

Harville, D.A. and Mee, R.W. 1984. A mixed model procedure for analysing ordered categorical data. *Biometrics* 40:393–408.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under selection model. *Biometrics* 31:423–447.

Henderson, C.R. 1984. *Applications of linear models in animal breeding*. University of Guelph, Guelph, Canada.

Henderson, C.R. 1986. Recent developments in variance and covariance estimation. *Journal of Animal Science* 63:208–216.

Henderson, C.R. 1988. Theoretical bias and computational methods for a number of different animal models. *Journal of Dairy Science* 71:1–16.

Hill, W.G. and Meyer, K. 1988. Developments in methods for breeding value and parameter estimation in livestock. Occasional Publication. *British Society of Animal Production* 12:81–95.

Huisman, A.E., Veerkamp, R.F. and van Arendonk, J.A.M. 2002. Genetic parameters for various random regression models to describe the weight data of pigs. *Journal of Animal Science* 80:575–582.

Jansen, G.B. 1990. Large-scale application of animal models for genetic evaluation of dairy cattle. *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production*, Edinburgh, Scotland. Volume 14:58–61.

Kotzé A. and Muller, G.H. 1994. Genetic relationship in South African cattle breeds. In: *Proceedings of the 5th world congress on genetics applied to livestock production, Guelph, Canada*. University of Guelph, Guelph, Ontario, Canada. Volume 21: 413–416.

Küttner, K. and Nitter, G. 1997. Effects of mating structure in purebred populations on the estimation of crossbreeding parameters. *Journal of Animal Breeding and Genetics* 114:275–288.

Lewis, R.M., and Brotherstone, S. 2002. A genetic evaluation of growth in sheep using random regression techniques. *Animal Science* 74:63–70.

Marshall, K., Quiros-Campos, C., Van Der Werf, J. H. J. & Kinghorn, B. 2009. Marker-based selection within small-holder productions systems in developing countries. Livestock Production Science, in print.

Mathur, P.K. and Horst, P. 1994. Methods for evaluating genotype-environment interactions illustrated by laying hens. *Journal of Breeding and Genetics* 111(4):265–288.

Meijering A. and Gianola D. 1985. Observations on sire evaluation with categorical data using heteroscedastic mixed linear models. Journal of Dairy Science 68:1226--1232

Meyer, K. 1987. Estimates of variances due to sire × herd interactions and environmental covariances between paternal half-sibs for first lactation dairy Production. *Livestock Production Science* 17:95–115.

Meyer, K. 1989. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genetic Selection Evolution* 21:317–340.

Meyer, K. 1998. Modelling `repeated' records: Covariance functions and random regression models to analyse animal breeding data. *Proceedings of the 6th World Congress on Genetics*

Applied to Livestock Production 25:517–520. University of New England, Armidale, Australia.

Meyer, K. and Hill, W.G. 1997. Estimation of genetic and phenotypic covariance functions for logitudonal or 'repeated' records by restricted maximum likelihood. *Livestock Production Science* 47:185–200.

Morrison, D.F. 1976. *Multivariate statistical methods*. McGraw-Hill, New York, USA.

Mrode, R.A. 2005. *Linear models for the prediction of animal breeding values*. 2nd Edition. CAB International, London, UK.

Nei, M. 1972 Genetic distances between populations. The American Naturalist.106, 282-292.

Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York, USA.

Nobre, P.R.C., Misztal, I., Tsuruta, S., Bertrand, J.K., Silva, L.O.C. and Lopes, P.S. 2003. Genetic evaluation of growth in Nellore cattle by multiple- trait and random regression models. *Journal of Animal Science* 81:927–932.

Ojango, J.M.K and Pollot, G.E. 2002. The relationship between Holstein bull breeding values for milk yield derived in both the UK and Kenya. *Livestock Production Science* 74:1–12.

Okomo, M.A. 1997. *Characterisation of the genetic diversity of East African cattle breeds using microsatellite DNA markers*. MSc thesis, University of Nairobi, Kenya.

Ota, T. 1993. *DISPAN: Genetic distance and phylogenetic analysis*. Pennsylvania State University, Pennsylvania, USA.

Park, S.D.E. 2001. *Trypanotolerance in West African cattle and the population genetic effects of selection*. PhD thesis, University of Dublin, Dublin, Ireland.

Patterson, H.D. and Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* 58:545–554.

Piry S, Alapetite A, Cornuet, J.-M., Paetkau D, Baudouin, L., Estoup, A. 2004. GeneClass2: A Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity* 95:536–539.

Pritchard, J. K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Ptak, E. and Schaeffer, L.R. 1993. Use of test day yields for genetic evaluation of dairy sires and cows. *Livestock Production Science* 34:23–34.

Raymond, M. and Rousset, F. 1995a. An exact test for population differentiation. *Evolution* 49:1280–1283.

Raymond, M. and Rousset, F. 1995b. GENEPOP-population genetics software for exact tests and ecumenicism. *Journal of Heredity* 86:248–249.

Saitou, N. and Nei, M. 1987. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406–425.

Schaeffer, L.R. 1991. *Class notes-Linear models and variance component estimation*. University of Guelph, Guelph, Canada.

Schaeffer, L.R. 2004. Application of random regression models in animal breeding. *Livestock Production Science* 86: 35–45.

Searle, S.R. 1971. *Linear models*. John Wiley and Sons, Inc., New York, New York, USA.

Snedecor, G.W. and Cochran, W.G. 1980. *Statistical methods*. 7th Ed. Iowa State University Press, Iowa, USA.

Stanton, T.L., Jones, L.R., Everett, R.W. and Kachman, S.D. 1992. Estimating milk, fat, and protein lactation curves with a test day model. *Journal of Dairy Science* 75:1691–1700.

Strandberg, E. 2006. Analysis of genotype by environment interaction using random regression models. In*: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production,* August 13–18, 2006, Belo Horizonte, MG, Brazil.

Swalve, H.H. 1998. Use of test day records for genetic evaluation. In*: Proceedings of the 6th world congress on genetics applied to livestock production*. University of New England, Armidale, Australia. pp. 295–301.

Swan, A.A. and Kinghorn, B.P. 1992. Evaluation and exploitation of crossbreeding in dairy cattle. *Journal of Dairy Science* 75:624–639.

Takezaki, N. and Nei, M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* 144:189–399.

van der Werf, J.H.J., Goddard, M.E. and Meyer, K. 1998. The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. *Journal of Dairy Science* 81: 3300–3308.

van Raden, P.M. 1997. Lactation yields and accuracies computed from test day yields and (co)variances by best prediction. *Journal of Dairy Science* 80:3015–3022.

Wiggans, G.R. and Goddard, M.E. 1996. A computationally feasible test day model with separate first and later lactation genetic effects. *Proceedings of the New Zealand Society of Animal Production* 56:19–21.

10 Related literature

Afifi, A.A. and Clark, V. 1990. *Computer-aided multivariate analysis*. 2nd ed. Chapman & Hall, New York, USA. pp. 371–393 and pp. 429–461.

Anderson, L. 2001. Genetic dissection of phenotypic diversity in farm animals. *Nature Review Genetics* 2:130–138.

Anderson, S.M., Mao, I.L. and Gill, J.L. 1989. Effect of frequency and spacing of sampling on accuracy and precision of estimating total lactation milk yield and characteristics of the lactation curve. *Journal of Dairy Science* 72:2387–2394.

Avise, J.C. 1994. *Molecular markers, natural history and evolution*. Chapman and Hall Publishers, New York, USA. 511 pp.

Bovenhuis, H. and Meuwissen, T. 1997. *Detection and mapping of quantitative trait loci*. Course compendium (16–20 June 1997). Centre for Genetic Improvement of Livestock, University of Guelph, Canada.

Brown, J.E., Brown, C.J. and Butts, W.T. 1973. Evaluating relationships among immature measures of size, shape and performance of beef bulls. I. Principal components as measures of size and shape in young Hereford and Angus bulls. *Journal of Animal Science* 36:1010–1020.

Casley, D.J. and Kumar, K. 1989. *The collection, analysis and use of monitoring and evaluation data*. A joint study of the World Bank, International Fund for Agricultural Development and FAO. Johns Hopkins University Press, Baltimore, Maryland, USA.

Clarke, B.E. and Kinghorn, B.P. 1997. A method to test algorithms for incorporating genetic marker data in BLUP. In: *Proceedings of the 12th conference of the Association for the Advancement of Animal Breeding and Genetics - Part I*. Dubbo, Australia, 6–10 April 1997. pp. 213–216.

Clarke, B.E., van Arendonk, J.A.M. and Kinghorn, B.P. 1997. Analysis of linkage between genetic markers and QTL using REML: The effects of selection on parameter estimates. In: *Proceedings of the 12th Conference of the Association for the Advancement of Animal Breeding and Genetics - Part I*. Dubbo, Australia, 6–10 April 1997. pp. 208–212.

Duchateau, L., Jansen, P. and Rowlands, G.L. 1998. *Linear mixed model: An introduction with applications in veterinary research*. ILRI (International Livestock Research Institute), Nairobi, Kenya.

Gill, J.L. and Hafs, H.D. 1971. Analysis of repeated measurements of animals. *Journal of Animal Science* 33:331.

Graser, H-U., Smith, S.P. and Tier, B. 1987. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *Journal of Animal Science* 64:1362–1370.

Jamrozik, J. and Schaeffer, L.R. 1997. Estimates of genetic parameters for a test day model with random regression for yield traits of first lactation Holsteins. *Journal of Dairy Science* 80:762–770.

Kennedy, B. 1991. Class notes-Linear models and variance component estimation. University of Guelph, Guelph, Canada.

Kettunen, A., Mantysaari, E.A. Stranden, I. and Poso, J. 1998. Estimation of genetic parameters for first lactation test day production using random regression models. In: *Proceedings of the 6th world congress on genetics applied to livestock production*. Armidale, Australia, University of New England, Armidale, Australia. pp. 307–400.

Krzanowski, W.J. 1988. *Principles of multivariate analysis: A users perspective*. Oxford Statistical Science Series. Oxford University Press, Oxford, UK.

Loftus, R.T., Ertugrul, O., Harba, A.H., El-Boyd, M.A.A., MacHugh, D.E., Park, S.D.E. and Bradley, D.G. 1999. A microsatellite survey of cattle from a centre of origin: The Near East. *Molecular Ecology* 8:2015–2020.

Luikart, G. and England, P.R. 1999. Statistical analysis of microsatellite DNA data. *Trends in Ecology and Evolution* 7:253–256.

Lukibisi, F.B. 2000. Statistical analysis of repeated measures: Livestock experimental data. In: Sustaining Animal Production into the 21st Century-Proceedings of the Animal Production Society of Kenya Symposium, 8–9 March 2000, Nairobi, Kenya. pp. 93–104.

MacHugh, D.E., Shriver, M.D., Loftus, R.T., Cunningham, P. and Bradley, D.G. 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* 146**:**1071–1086.

Meyer, K. 1991. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genetic Selection Evolution* 23:67–83.

Meyer, K. and Smith, S.P. 1996. Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genetic Selection Evolution* 28:23–49.

Olori, V.E., Hill, W.G., McGuirk, B.J. and Brotherstone, S. 1999. Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. *Livestock Production Science* 61:53–63.

Quaas, R.L. and Pollak, E.J. 1980. Mixed model methodology for farm and ranch beef cattle testing programs. *Journal of Animal Science* 51:1277–1287.

Searle, S.R. 1982. *Matrix algebra useful for statistics*. John Wiley and Sons, Inc., New York, New York, USA.

Steel, R.G.D. and Torrie, J.H. 1980. *Principles and procedures of statistics: A biometrical approach*. 2nd ed. McGraw-Hill Book Company, New York, USA. 481 pp.

Swalve, H.H. 1995. The effect of test day models on estimation of genetic parameters and breeding values for dairy yield traits. *Journal of Dairy Science* 78:929-938.

Wright, S. 1932. General group and special size factors. *Genetics* 17:603-619.

Yates, F. 1981. Sampling methods for censuses and surveys. 4th Ed. Charles Griffin & Co. Ltd. London, UK.

11 Websites

The web pages were all accessed in 2009.

11.1 Data bases

Gene maps, databases etc. http://bos.cvm.tamu.edu/bovarkdb.html

11.2 Software

Programmes for estimation of variance/covariance components and/ or prediction of breeding values: (see software)

ASREML: http://www.genstat.com/products/asreml

VCE: http://www.tzv.fal.de/institut/genetik/vce4/vce4.html

PEST: http://www.tzv.fal.de/~eg

WOMBAT: http://agbu.une.edu.au/~kmeyer/wombat.html

Programmes for estimation of crossbreeding effects:

CBE - Crossbreeding Effects: http://www.boku.ac.at/nuwi/software/softcbe.htm

Programmes for measuring genetic diversity based on genetic markers:

Analysis of Molecular Variance,

AMOVA: http://www.bioss.ac.uk/smart/unix/mamova/slides/frames.htm

ARLEQUIN: http://anthro.unige.ch/arlequin

DISPAN: http://www.bio.psu.edu/People/Faculty/Nei/Lab/Programs.html

Programmes for QTL mapping:

QTL Express http://qtl.cap.ed.ac.uk/ (Regression mapping; Interval mapping, inbred and outbred populations)

QTL cartographer http://statgen.ncsu.edu/qtlcart/WQTLCart.htm (Maximum Likelihood mapping; Composite interval mapping in experimental populations)

11.3 Courses and course notes

Schaeffer's Note shop with course notes for animal models, quantitative genetics and methodology in animal breeding: http://www.aps.uoguelph.ca/~lrs/Animals/

Course notes on gene mapping and QTL in breeding:

http://www-personal.une.edu.au/~jvanderw/aabc_materialsp3.htm

http://www-personal.une.edu.au/~jvanderw/Models_for_QTL_analysis.pdf

11.4 Miscellaneous

Alphabetic list of genetic analysis software (population genetics software and linkage analysis) http://linkage.rockefeller.edu/soft/