Module 4
Animal Genetics Training Resource (CD-ROM)
Version 1 (2003)
ILRI-SLU

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

Ntombizakhe Mpofu[1] and Lena Andersson-Eklund[2]

1. International Livestock Research Institute (ILRI),
P.O. Box 5689, Addis Ababa, Ethiopia (sections 1–6)

2. Swedish University of Agricultural Sciences (SLU), Dept. of Animal Breeding and
Genetics, P.O. Box 7023, SE-75007 Uppsala, Sweden (sections 1 and 7)

The purpose of the present module is to give a review of the most commonly used quantitative methods in the area of animal breeding and animal genetic resources. The module primarily addresses scientists in the area of animal genetic resources in developing countries, including both faculty in universities/colleges and staff in research institutions. The core text includes links [blue] and references to other parts of the AnGR Training Resources (CD-ROM), such as exercises, module texts, compendia and case studies. There are also links [burgundy] and references to literature and websites.

Contents

1 Quantitative methods—Important tools for AnGR

1.1 Statistics to separate genetic and environmental effects

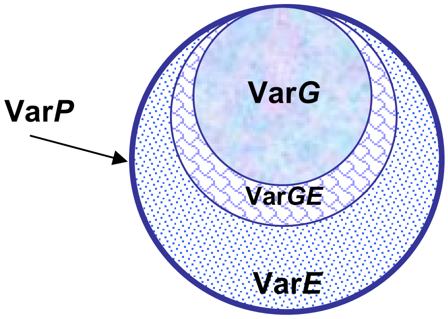
The observed characteristics of animals, which is the phenotype of the animal is affected both by genetic and environmental factors. The genetic factors are due to random sample of genes received from the two parental gametes, whereas, the environmental factors include influences by climate, nutrition, health and management. Genetic analyses in the field of AnGR most often aim at separating genetic and environmental effects. For that purpose, we need a mathematical model 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 (1989) presented as:

$$\mathrm{P = G + E + G \times E}$$

It is important that we consider the specific combination effects between genotype and environment, and therefore, include an interaction term G×E in the model [Module 2, Section 3.4].

When we consider breeds or populations of animals, P and G are expressed as deviations from the population mean and we can describe and illustrate the variation within the population as:

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



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. It is therefore important to determine to what extent a phenotype is influenced by genetic effects (i.e. the extent a trait is heritable) and then design breeding programmes accordingly. With a single individual, it is not possible to separate the effects of genetic and environmental factors and 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 2–5 below provide us with powerful tools for analysing and handling quantitative variation in practical breeding

1.2 Use of statistics to estimate genetic diversity

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 [Module 2, Section 3]. The Section also describes molecular genetics techniques that can be used to collect data for studies of genetic diversity. In the present module, some quantitative methods that are used to analyse molecular data for measuring genetic diversity in populations [Section 6.1, this Module] and genetic relationships between populations [Section 6.2, this Module] are reviewed.

1.3 Statistics for genetic dissection of traits

Currently, a lot of research effort is being put into studies of molecular genetic background of traits in livestock. For this purpose, quantitative methods for analysing phenotypic and molecular genetic data have been developed. The majority of production, functional and health traits are the consequences of complex physiological systems in the animal. They are thus influenced by a large number of genetic and environmental factors. The animals' phenotypes do not fall into discrete classes, but show continuous variation. As the number of genes and gene interactions influencing each trait is expected to be very large, it is evident that the genetic basis for such quantitative traits can neither be fully clarified nor considered in full detail, like what 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. 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.

As compared to studies of phenotypic distributions of traits, studies of linkage 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 localise QTL for many economically important traits in livestock species. For the detection of QTL, it is essential to make use of efficient statistical methods some of which are reviewed in (see Section 7:1) below.

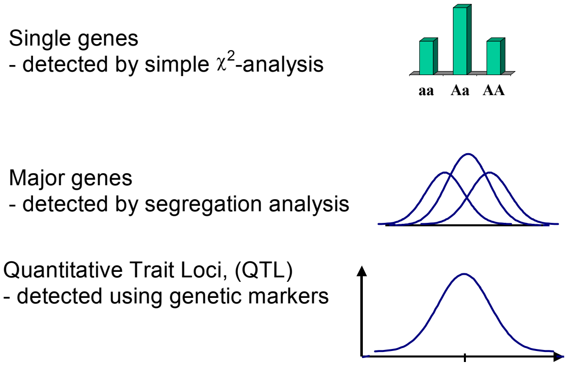

Single genes
- detected by simple χ^2-analysis
aa
Aa
AA
Major genes
- detected by segregation analysis
Quantitative Trait Loci, (QTL)
- detected using genetic markers

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

2 Understanding your data

The first step in genetic studies is to collect data from a number of animals. There are two usual sources of data used in animal breeding. Research scientists set up experiments and collect data from experimental animals. 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 as poorly designed experiments do not produce useful data. Data must be amenable to a statistical analysis from which we can draw inferences or can predict future observations. However, for most breeding works, large data sets are required to get reliable estimates of phenotypic, genetic and environmental parameters. Costs usually prohibit the setting up and running of large experiments to collect the required data. Data is, therefore, often obtained from farms (field records) through livestock recording schemes (Module 3, Section 3.2).

Data can be subjective (e.g. body score) or objective (e.g. body weight recorded on a scale). In both cases they can be analysed. The first step in data analysis is to check the data for possible errors in data recording or computer entry. Errors can then be corrected or records with error deleted. It is also important at this stage to understand the data structure and the patterns displayed in the data in order to decide how best to conduct the statistical analysis [Biometrics example 1]. The distribution of animals by different classification (e.g. age, sex) can be determined and mean, median and range for each factor or classification variable summarised. 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 equal number of observations per subclass can end up having different numbers of observations after data editing. Data with unequal number of observations per subclass are known as unbalanced data and 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 model that is used to describe the data best judges the quality of any statistical analysis. An appropriate model can only be chosen when one understands the data.

3 Constructing a model

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. But 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 accurately enough. Whenever an operational model (instead of an ideal one) is used, it is recommended that the ideal model is outlined and reasons for not using and problems likely to arise from not using it are given. The statistical models commonly used in animal breeding are linear models, with the set of factors being assumed to additively affect the observations. This does not mean that non-linear models are not important in animal breeding. Rather, the choice of linear models has been influenced by the traits studied and their importance (Schaeffer 1991).

As an 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. These factors or effects can be discrete or continuous. Discrete factors have distinct levels. For example, 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 are fitting a straight line. The slope of this line is known as a regression coefficient [Biometric 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 [Biometric example 2].

A co-variable is a factor known to affect a performance trait but is not one of primary interest for the outcome of the statistical model. By including co-variables in the model, however, adjustments are made to the mean values of the primary factors of interest and corrections for corresponding variation in the mean values of the co-variables themselves are made. When there is a significant relationship between the trait being analysed and a co-variable, a proportion of the natural variation among animals is explained, and this in turn improves the precision of comparison between mean values of primary interest [Biometrics example 2].

An effect (factor) can be fixed or it can be random. When we think of a factor as fixed effect, we infer that any conclusion drawn about the estimated mean for the trait applies only to the study itself. 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 random and any inferences made from the study are generalised 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.

A model comprises three parts: (1) the equation which describes the factors (effects) and their levels, (2) estimates of means and variances 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 both fixed and random effects. Thus, according to Searle (1971); 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

Each of these models involves various assumptions. For example, residuals should normally be distributed and each observation should be randomly and independently obtained. Repeated measures on an animal can cause some difficulties because adjacent observations may be closely correlated than those further apart. Statistical procedures are generally fairly robust and slight departures from normality can be ignored. When data are clearly not distributed normally some kind of transformation (e.g. a logarithmic or square root transformation) of the data can be considered.

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 value of a trait being evaluated can be influenced by many factors. For example, dairy data can include records of thousands of herds made over many years, some information can be missing for some herds or years, cows within the herd can be of various genotypes and/or age, 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. Different computational techniques have been developed to deal with such data — e.g. aborbing a factor to reduce the size of the matrix being inverted, solving the model iteratively by including certain covariables or secondary factors in a preliminary model 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 normal 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 most cases, however, it is advisable to use logistic regression approach (Krzanowski 1988), which is a form of a generalised linear model that caters for binomially distributed trait.

4 Separating non-genetic and genetic effects

4.1 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 levels, others as co-variables. Some fixed effects influence data but are in themselves of little interest. Data is corrected and no estimations are done for these. However, there are some effects (e.g. trends in the year or differences in sexes) whose estimates may be of interest. For these estimations are done.

Estimates for the parameters of the model can be obtained and the most common criterion of estimating these parameters is least squares (Searle 1971). Once the parameters have been estimated, tests can be done to find out whether the factors included in the model account for significant variation in the production trait measured. For example, for the classification models, estimation of linear functions and testing hypotheses related to those functions is done. For the effect of season of calving, for example, we may want to test that milk yield for cows calving in winter differs from those calving in 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 that seasonal differences are not important for milk yield [see Biometrics example 2]. In this example, least squares analysis fitting fixed effects (discrete and continuous) is illustrated. The steps followed are: calculating 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. Nowadays there are procedures that can estimate parameters for environmental factors, adjust the data for these factors and estimate genetic effects simultaneously. These are recommended procedures but where the required software is unavailable the important environmental parameters need to be determined by general least squares methods and the raw data corrected for these effects before proceeding to estimate genetic parameters.

4.2 Estimating genetic effects

Often, rather than estimating specific differences between treatments, we may be interested in estimating variances due to effects – see Section 3. For example, in milk production, we may not be interested in estimating differences between dams but rather in estimating the variation among them as an estimate of the variation from which they were sampled. The dams can be considered as random effects and the data can be analysed according to a random effects model [Biometric example 3].

4.2.1 Variance component estimation

The data described in 4.1 above is used to estimate genetic and environmental variances which are needed in calculating genetic parameters (e.g. heritability) and in 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) can be analysed fitting a sire model, and co-variances generated by these relationships provide the information required for estimation of additive genetic variance. However, linear models containing both genetic and environmental effects for each animal (animal model) have become more popular.

Statistical procedures for variance component estimation include analysis of variance (ANOVA) type methods. These require equal number of observations in all subclasses– balanced design). With unbalanced data, the ANOVA method for mixed models leads to biased estimators of variance components. Henderson developed Henderson’s methods 1, 2 and 3, which were modelled after balanced ANOVA (Henderson 1984). These methods had weaknesses that other methods developed later tried to address. For example, Henderson’s method 2 first uses data to estimate the fixed effects of the model. Data are then adjusted by these estimators, and the variance components are estimated from the adjusted data. Henderson’s method 2 cannot, therefore, be used when there are interactions between fixed effects and random effects.

The widely applied methods in variance component estimation are maximum likelihood (ML) procedures. These procedures estimate the fixed effects and variance components simultaneously. Animal breeders are more and more confronted with data sets that have arisen from either selection experiments or from farm testing in which selection has been practised. If lack of records is as a result of selection based on some criterion that is correlated to trait(s) under analysis, the resultant estimates are likely to be biased by selection. ML estimation procedures utilise all records available and, can account for selection (Harville 1977). A modified ML procedure, i.e. restricted maximum likelihood (REML) which accounts for the loss in degrees of freedom due to fixed effects in the model of analysis (Patterson and Thompson 1971), has become the preferred method of analysis in animal breeding, not least for its ability to reduce selection bias. Graser et al. (1987) suggested a derivative-free method of estimating variance components, based on sequentially calculating the likelihood. The derivative-free approach provides a robust, flexible and powerful alternative to derivative-based REML algorithms. Its application for animal models, in which additional random effects, for instance animal's maternal genetic effects or common environmental effects are included, for the univariate case was described by Meyer (1989) and its extension to multivariate analysis is given in Meyer (1991). More recently an average information algorithm giving improved rates of convergence has been implemented (Meyer and Smith 1996). In general, 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) and DFREML (developed by Karen Meyer) [Web pages, Section 9, this Module].

4.2.2 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, in many cases, financial and management constraints result in poor data sets.

Realised 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 realised 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 best linear unbiased predictor (BLUP) (Henderson 1984). The differences between BP, BLP and BLUP are subtle but yet important statistically [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 [DFREML] (seeWeb pages, Section 9, this Module). Various sources of information can be used to calculate BLUP breeding values. For example, when sire information is used, we have a sire model: and if sire and dam information is used, we have a sire and dam model. With BLUP, we can also do what the farmers have always wanted, i.e. combine all the pedigree information and performance details for all relatives in the evaluation of an individual (animal model) [Computer exercises, BLUP]. BLUP helps to remove some of the biases, such as selective mating, and takes account of genetic trend. If there are sufficient connections between herds, as is usually the case with the use of AI, selection on BLUP values can be done on a breed (rather than herd) basis [Manual exercises: Selection index].

Breeding values can be estimated for each trait of interest. Hence, selection decisions can be made based on that trait. However, in livestock production, farmers are usually interested in improving more than one trait at a go. For such situations, information on several traits can be used to construct a multiple trait breeding value (by BLUP or selection index) [Manual exercises: Selection index] given that genetic and environmental correlations between traits can be estimated.

Test day models in dairy production
Genetic evaluations for dairy cattle in many countries are obtained by analysing 305-day (or equivalent cumulative yield records) predicted from a few test day yields. The 305-day yields predicted from monthly test day records can be inaccurate and biased. 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. Genetic evaluations based directly on test day records can overcome the need to predict 305-day yields or project incomplete lactations. They can also facilitate a cheaper and more flexible recording scheme. They, therefore, offer an opportunity to improve the genetic evaluation of dairy cattle in tropical production situations where there are limited infrastructures to support more sophisticated or detailed recording systems, often resulting in too small data sizes to allow for accurate genetic evaluation of bulls as production conditions are constrained by environment and resources (Swalve 1998). Many studies have been done to explore the potential of statistical and computing techniques that allow a direct and more efficient utilisation for genetic evaluation of all available records (e.g. Swalve 1995 and 1998; Jamrozik and Schaeffer 1997).

4.2.3 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 progenies 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 interaction of genotype and environment (G x E) would reduce potential benefits from a strategy based entirely on continuous importation of superior germplasm from elsewhere [CS 1.16 by Mpofu]. The resultant G x E may affect ranking of genotypes and, indeed, individuals depending on the country and production system (Ojango and Pollot 2002). The magnitudes, in absolute or relative terms, of the genetic, residual and phenotypic variances obtained from populations raised and recorded in countries with different environmental constraints are often also different.

Methods of estimating GxE are reviewed in Mathur and Horst (1994) and Chagunda (2000). They 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 environmental factors, genetic factors 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.

Correlation of breeding values
This procedure is used when the same sires have progeny in the two environments. The product-moment correlation between breeding values of sires estimated in two environments gives an estimate of the genetic correlation between the environments. When calculating the correlation between proofs, the proofs made in the two countries need not be weighted by number of daughters when the method used to calculate proofs has already considered the amount of information going into the proof (Mpofu 1992; Ojango and Pollot 2002).

Estimation through selection in two environments
GxE can also be determined indirectly from direct and correlated response to selection (Falconer 1989). 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 direct response possible through selection in environment Y. The ratio of correlated response and direct response is computed and used to calculate GxE. This method, although it is likely to give a reliable measure of GxE, can only be applied after selection has been practised.

4.2.4 Estimating heterotic effects

Crossbreeding is a widely accepted livestock production practice in developing countries. The basis of systematic crossbreeding can broadly be classified into additive and non-additive. The additive component is that which is due to the averaging of 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 crossbreeding. 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.

The performance of crosses can be predicted using estimates of genetic parameters from crossbreeding experiments. Models for estimating crossbreeding parameters based on two-locus factorial model of gene effects have been developed earlier by Dickerson (1973) and lately by Küttner and Nitter (1997). A case study by Kahi [CS 1.5 by Kahi] illustrates an example of data analysis for estimating crossbreeding parameters for milk production traits under the humid coastal regions of East Africa, while another by Aboagye [CS 1.9 by Aboagye et al.] gives such parameters for milk production, reproductive, growth and carcass traits in cattle under the humid West African tropical conditions. Software such as CBE (Crossbreeding effects) are also available that be used to estimate crossbreeding effects from a larger variety of data structures or experimental designs [Section 9, this Module].

5 Designing and implementing on-farm surveys of livestock breeds

In livestock production, population censuses are carried out at given intervals. Normally such censuses are conducted to estimate the number of animals by species (e.g. number of cattle, sheep or goats) or by enterprise (e.g. number of dairy cattle) and the information collected is used for administrative or planning purposes. In most cases, the censuses do not record breed types. However, developments in livestock improvement and reproductive physiology have resulted in some breeds being more popular than others. The unpopular breeds then become threatened by extinction. It has become necessary and increasingly so to undertake breed surveys so as to determine the status of different breeds in a country and then use this information to develop breed improvement and conservation strategies.

5.1 Designing on-farm surveys of livestock breeds

The first step is to decide what type of survey (random, purposive or representative etc.) is to be undertaken and then the size of the population to be surveyed. The whole population (complete census) or samples of the population can be surveyed. Where a sample is to be surveyed, a decision on the proportion of the farming community or households to be surveyed needs to be made. The size of the sample needs to be large to allow population values derived from the sample to be estimated with adequate precision. At the same time, costs in collecting data need to be considered as these tend to increase with an increasing sample size. Statistical methods that allow one to determine the sample size necessary to estimate population values with a required level of precision are available. Different sampling designs are available from simple random sampling to those using stratified and clustering techniques.

Data are usually collected using questionnaire forms. The questions need to be designed in a way that allows accurate, unambiguous answers to be given, which can provide data for sound statistical analysis.

Pre-testing of a questionnaire on a small number of farms or households is an essential and a very useful way of evaluating the suitability of the questionnaire and the level of detail that is possible to obtain from the interviewees. It is also important that the survey is designed taking into consideration the statistical analysis that will be carried out, again the aim being to collect sufficient information for every subclass. . 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 it is important to ensure that:

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

These can be obtained during pre-survey visits, otherwise it would be almost be impossible to be able to estimate, or project how the variances estimates for the mean population, regional values would be, hence how best to achieve high accuracy and precision at the same time [see Module 2, Section 2].

5.2 Implementing on-farm surveys

In implementing on-farm surveys, many things should be considered and undertaken. These include 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); visiting time and where to interview respondents (in the homestead, on grazing fields) and who the respondents should be (household heads, children, employees). It is important to note that 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 are best given by the family members who actually do the milking, although the norm may be for the household head to respond to such questions or the entire questionnaire.

Although breed descriptor charts and guidelines on animal phenotypic characteristics, such as those developed by ILR and used for the Oromiya-ILRI Livestock Breed Survey (2001), may be available to assist enumerators and questionnaire administrators in making on-farm survey decisions, occasional use of photographs to capture whole herds, while in pens, kraals or grazing, greatly help to counter check 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 respondent, seem to giving pre-planned answers or non-plausible ones.

It is always disastrous to begin a survey, with incomplete plans and inadequate resources in place. On the other hand, sequential survey can accomplish a lot, whenever foreseen financial and logistic inadequacies are taken into account and, are thus included in the technical planning (design) process.

After entering the data into a computer and checking for errors and verification, the analysis of survey data starts with investigating the patterns in the data [Biometrics example 1]. This involves tabulating the information and calculating simple statistics. The data can then be analysed as one set or divided into subsets for analysis. There are methods that can be used to determine the size of these sub-samples. Statistical models can then be developed to test different hypotheses suggested by the preliminary analyses. Any type of models described earlier, e.g. analysis of variance or regression analysis, can be used.

Sample estimates can be used to estimate population values. Formulae are available for calculating population mean and their standard errors for stratified or cluster sampling designs.

6 Measuring genetic diversity in populations

6.1 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 breed 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.

6.1.1 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 either genetic isolation, or historical population bottlenecks, or founder effects. A high MNA implies great allelic diversity, which could have been influenced by crossbreeding or admixture. Bar charts can be created for individual breeds to show variability in allelic distributions at loci.

6.1.2 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.

6.1.3 Variation in genotype frequencies

Variability between breeds can be measured using the observed genotypes at each locus and between pair 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.

6.1.4 Testing for Hardy-Weinberg equilibrium

The relationship between gene frequencies and genotype frequencies is of great importance because most deductions about populations and quantitative genetics depend on it. A population is said to be in Hard-Weinberg equilibrium 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. A Hardy-Weinberg test is performed to assess the genetic structure within an individual breed, i.e. to check whether the gene frequencies significantly differ from the expected ones. The data required are gene and genotype frequencies and the size of sample population at each locus.

The deviation from Hardy-Weinberg equilibrium can be tested using any one of the following three methods.

(a) The Chi-square test which 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 Hardy-Weinberg equilibrium. In addition, for loci or population combinations with five or more alleles, a Markov chain algorithm is used to obtain unbiased estimate of the exact probability of being wrong in rejecting Hardy-Weinberg Equilibrium. The GENEPOP package (Raymond and Rousset. 1995) can be used to do the test.

6.1.5 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 or due to the mixed nature of the breeds or due 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 DISPAN computer program (Ota 1993) can be used to estimate expected heterozygosity.

6.1.6 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. It has been showed that there was no significant difference between average inbreeding coefficient values estimated from pedigree data and biochemical data (Avise 1994).

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 Hardy-Weinberg equilibrium. A positive inbreeding coefficient value reflects the existence of small number of heterozygote genotypes and excess of homozygote genotypes. A negative value indicates the occurrence of heterozygote genotypes at higher proportion than the homozygote genotypes.

6.1.7 Genetic differentiation

Population differentiation can be assessed by determining whether allelic composition is independent of population assignment (Raymond and Rousset 1995). The statistical test is based on analysis of contingency tables using a Markov Chain procedure to derive unbiased estimate of the exact probability in being wrong in rejecting the null hypothesis, i.e. allelic composition is independent of population assignment (no differentiation). The test is performed for pairwise inter-population comparisons on contingency tables containing data from each of the microsatellite loci studied. The GENEPOP package can be used.

6.1.8 Analysis of gene flow and genetic admixture

(a) Use of diagnostic allele

Diagnostic alleles are alleles that are unique to certain breeds, e.g. allele 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 rational 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 indicine 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 solution for each individual admixture proportion with associated standard errors. It also produces correlation coefficients for the weighted least squares solutions that gives 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)

6.1.9 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 or (ii) chance in small populations (e.g. endangered breeds) or (iii) selection favouring one combination of alleles over another or (iv) the close association between markers in the same linkage group (Falconer 1989). A test can be carried out to check for the existence of the association between markers studied. The null hypothesis for LDE test is that all the genotypes at one locus are independent from those at another locus. The GENEPOP program (Raymond and Rousset 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.

6.1.10 Distribution of genetic diversity

The distribution of genetic diversity within and between populations is generally done through an analysis of molecular variance (AMOVA), which is essentially an analysis of variance (ANOVA). This procedure uses information from both the estimated divergence between haplotypes and the frequency at which each is represented in a population grouping. The first step is to create a distance matrix between samples in order to measure the genetic structure of the population from which samples are drawn. Then variance components are estimated.

AMOVA separates and tests tiers of genetic diversity: among groups of populations, among populations within groups and among the individuals within a population. The variance components from the analysis are used to estimate phi (Φ) statistics that are similar to F statistics. The size of the Φ statistic gives an indication of changes in diversity over time.

Software for AMOVA analysis (called AMOVA) was developed by staff of the University of Geneva and is available freely on the Internet. A much more sophisticated program called ARLEQUIN was later developed by the same group and can handle data from RFLP, DNA sequences, microsatellites as well as standard multi-locus or allele data [see Section 9, this Module].

6.2. 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-Adhiambo]; [CS 1.11 by Gwakisa].

6.2.1 Cluster analysis

Clustering is a technique for grouping individuals into unknown groups to demonstrate 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 prior to the analysis. In animal diversity studies, cluster analysis has been used in classifying breeds into groups on the basis of their genetic characteristics. Before clustering it is usually recommended to do some initial analysis. 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 often highly correlated. 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 in 1972, DA distance developed by Nei in 1983 and a genetic distance measure developed by Goldstein *et al.* in 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 and an example is DISPAN (Ota 1993) (see Section 9, 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 the ones 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 neighbour-joining method (NJ) (Saitou and Nei 1987). The operational taxonomic units (OTUs) in breeding are livestock populations or breeds. Therefore, the phylogenetic trees summarise evolutionary relationships among breeds or populations and categorise 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 the separation period between breeds [CS 1.10 by Okomo-Adhiambo]; [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, 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 very 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. Also, the underlying situation may suggest a natural number of clusters.

6.2.2 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 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 are 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 more concise summarisation of data 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 in order 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.

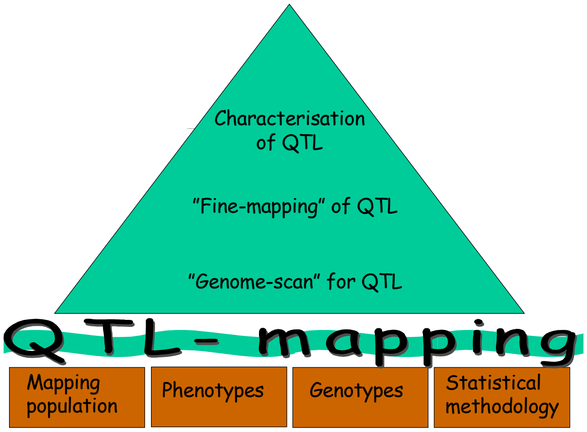
The PCA procedures in genetic studies were described by Cavalli-Sforza et al. (1994). In animal genetic diversity studies, PCs have been used to determine relationships among populations, supplementing relationships determined using phylogenetic analyses (e.g. Okomo 1997). PCs can be more convenient 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 trees.

7 Mapping quantitative trait loci (QTL)

7.1. Strategies for QTL analyses

The aim of QTL analyses is to detect, localise 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. In some cases, candidate genes for QTL are known based on information from other populations or other species. If there are known candidate genes, these 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. The genome scan can show in which chromosome segments 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 2) 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 [see Section 9, this Module].



Fugure 2. *The phases of QLT mapping*

A full genome scan for QTL includes five steps:

(i) *Choice of a mapping population*: In domestic animals, we can either use experimental crosses between divergent populations or large families within a population. Studies in designed crosses are powerful as they create high linkage disequilibrium, and large QTL effects segregate in, e.g. a backcross or intercross designs. However, such experiments are very costly for large animals and they do not give any direct answers of the segregation of QTL within the commercial populations of interest. Therefore, for large animals like cattle, mapping studies are usually performed in existing populations, within families or by selection of individuals with extreme phenotypes.

(ii) *Collection of phenotype data*: To get an acceptable power in the analyses the phenotypes have to be recorded on large numbers of animals. They can either be the same animals that are genotyped and/or offspring of the genotyped individuals (progeny testing).

(iii) *Genotyping:* Genetic maps, based on DNA markers, are available for many species. In livestock, short tandem repeats or microsatellites are currently the markers of choice as they are highly polymorphic and more than a thousand of them are available in most species. 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) as well as remaining genetic and environmental effects is used.

(v) *Drawing statistical inference from data*: The statistical testing for QTL is performed at marker loci (single marker analysis) or at marker loci as well as in intervals between markers (interval mapping). Multiple testing across the genome must be considered when setting significance thresholds. Parameters are estimated in the most likely positions for QTL by regression, ML, BLUP-based or MCMC (Markov Chain Monte Carlo) methods.

7.2 Why map QTL?

The detection and localisation 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 utilised to improve the efficiency of selection schemes (see Module 3, Section 4.7). 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.

8 References

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

Cavalli-Sforza L.L., Menozzi P. and Piazza A. 1994. *The history and geography of human genes*. Princeton University Press, Princeton, New Jersy, 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. (e*ds), Genetic microdifferentiation in man and other animals*.). Indian Anthropological Association, Delhi, India. pp. 171–180.

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.

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

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

Dunchateau 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.

Falconer G. 1989. *Introduction to quantitative genetics*. 3rd edition. Longmans, Harlow, UK.

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

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.

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.

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.

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

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.

Kettunene 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.

Kotze 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, 7–12 August 1994*. Volume 21. University of Guelph, Guelph, Ontario, Canada. pp. 413–416.

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

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.

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 200 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.

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.

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. 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.

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

Mpofu N. 1992. Genetic and economic evaluation of dairy cattle breeding strategies for Zimbabwe. PhD thesis, University of Guelph, Guelph, Canada.

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

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.

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.

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

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

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

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

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

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.

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

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

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

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, 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.

Weller J.I. 2001. *Quantitative trait loci analysis in animals*. CABI Publishing. London, UK. 304 pp.

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

9 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.

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, Mariland, USA.

10 Websites

The web pages were all accessed in July 2003.

10.1 Data bases

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

10.2 Software

Programmes for estimation of variance/covariance components and/or prediction of breeding values:

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

DFREML: http://agbu.une.edu.au/~kmeyer/dfreml.html

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

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:

Regression mapping; Interval mapping, inbred and outbred populations

QTL Express http://qtl.cap.ed.ac.uk/ Maximum Likelihood mapping; Composite interval mapping in experimental populations;

QTL cartographer http://statgen.ncsu.edu/qtlcart/WQTLCart.htm

10.3 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

10.4 Organisations and networks

Reeves J. C., J.R. Law, P. Donini, R.M.D. Koebner, and R.J. Cooke. Changes over time in the genetic diversity of UK cereal crops: http://apps3.fao.org/wiews/Prague/Paper12.htm

10.5 Miscellaneous

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