UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

SCHOOL OF MATHEMATICS

PROJECT WORK

APPLICATION OF LINEAR MIXED MODELS IN MICROARRAY
A PROJECT SUBMITTED TO THE SCHOOL OF MATHEMATICS IN PARTIAL FULFILLMENT
FOR A DEGREE OF MASTER OF SCIENCE IN BIOMETRY.

AUGUST 2007

by
Daniel Kimuyu Mwero
Supervised by
Dr. Thomas Achia
(University of Nairobi)

and
Dr. Etienne De Villiers
(International Livestock Research Institute - Nairobi, Kenya)

Declaration

I the undersigned declare that this project is my original work and to the best of my knowledge has not been presented for the award of a degree in any other University.

Daniel Kimuyu Mwero
Reg. No I/56/7419/2005

Signature..... Date.....05/09/07.....

Declaration By Supervisor

This project has been submitted for examination with my approval as supervisor

Dr. Thomas Achia
School of Mathematics,
University of Nairobi,
P.O. Box 30197 Nairobi,
KENYA.

Signature..... Date.....05/09/07.....

Acknowledgments.

I wish to express my sincere gratitude and appreciation to my supervisor, Dr. T.N.O Achia for his guidance, technical advice,patience and encouragement throughout the course of this work.
I also can't thank enough Dr. Etienne de Villiers for giving me this project, supervising me and providig all the financial support needed throughout the entire project. Many thanks to Dr. Rob Skilton for teaching me all the microarray techniques and for his patience especially during the bench work in the lab.
The invaluable support by Members of staff in the School of Mathematics through their guidance and constructive feedback during the whole project period is also highly appreciated. My classmates especially Thuranira, Mutua, Ng'etich, Waithaka and Mbunzi,thanks for your invaluable support.
The financial support afforded by Sophie will forever be appreciated.
Finally I would like to thank my family for the moral support and prayers during the whole course of this work.
God bless you all.

Dedicated:

To God,

My Parents Mr & Mrs J. Mwero,

My Siblings Julliet, James, Janet and John,

and

Sophie the love of my heart

for believing in me.

Contents

Abstract.

This project captures the problem of large microarray datasets and seeks to identify a statistical model of microarray hybridization intensity data that describes;differential regulation,sample-sample variability and measurement noise. It also shows how one can use the data model to analyse the microarray data and develop optimal methods for detecting differentially regulated peripheral blood leukocyte mRNA from cattle infected with Trypanosoma congolense using microarray in order to assay components of the immune and inflammatory responses and identify potential correlates of the pathology. We conclude by giving an insight into linear mixed effects models by analysing a data set from a cattle experiment that seeks to compare 'genome-wide' transcriptional responses in blood leukocytes following infection with species of Trypanosoma that differ in the severity of pathogenicity.

Chapter 1

Introduction

1.1 Background

Trypanosomosis, a disease/infection caused by the trypanosome parasite, in livestock is a disease caused by protozoan parasites of the genus Trypanosoma and is transmitted cyclically by the tsetse fly (Glossina spp). Trypanosomes are transmitted by the bite of the tsetse-fly and live in the bloodstream of infected host animals. The most important species of trypanosome that cause disease in livestock include Trypanosoma vivax, T. congolense and T. brucei brucei (Ikede, 1981), and are widely distributed in agroecological zones of Sub-Saharan Africa. Other species, T. simiae and T. evansi cause diseases in pigs (Leach and Roberts, 1965) and camels (Scot, 1973) respectively. African animal trypanosomosis has been described as a major obstacle to sustainable livestock production and food security, and an important factor of underdevelopment in sub-saharan Africa (Onyiah, 1997; Swallow, 2000). Trypanosomosis, equivalent to human sleeping sickness, has been described as the single most devastating disease in Africa in terms of poverty and loss of agricultural production. The estimated loss amounts to 5.5 billion dollars annually (Hursey, 2000). Furthermore, it is estimated that 50 million cattle are at risk of becoming infected with trypanosomosis leading to more than 3 million livestock deaths yearly, losses in calving, reduction in livestock numbers, drop in meat and milk off take and reduced work efficiency of draft animals and profitability of mixed farming (Budd, 1999; Hursey, 2000). Recent observations show that trypanosomosis is a major cause of culling of livestock and has impact on the physical condition of cattle at slaughter and consequently, the market value (Abenga *etal.*, 2002). Diseased cows become cachectic and fail to convert the food that they consume into products: no growth, no milk, no calves, and no draught power. Current trypanosomosis control relies on trypanocidal drugs, use of trypanotolerant cattle breeds and control of the tsetse fly vec-

tor. None of these methods have the full potential to work in the long-term control of the disease. Most heavily relied on are the trypanocidal drugs and this has lead to an increasing problem with resistance in the target organisms. These current methods of control for trypanosomosis are inadequate to prevent the enormous socioeconomic losses resulting from this disease and a vaccine has been viewed as the most desirable control option. Research towards a vaccine requires the identification of immune mechanisms involved in parasite and disease control or those responses that are associated with pathogenicity. The objective of this project is to employ a functional genomics approach using a bovine cDNA microarray platform to compare 'genome-wide' transcriptional responses in peripheral blood leukocytes (cells involved in the immune defense of the body) following infection with species of Trypanosoma that differ in the severity of pathogenicity. This will involve two groups of cattle which will be infected intravenously (I.V.) by needle with bloodstream forms of T. evansi and T. congolense. The cattle will be monitored comprehensively for up to 4 weeks. Peripheral blood leukocytes will be collected at specific stages of infection and gene expression will be measured on a 15,000 element bovine cDNA microarray using leukocyte RNA. Microarraying is a powerful functional genomics technology which permits the expression profiling of thousands of genes in parallel (Schena *etal.*1996). Typically, PCR-amplified cDNA's are printed at high density onto the surface of coated glass microscope slides (probe DNA) and hybridised with two fluorescently-labeled targets made from mRNA derived from the cell types of interest. One of the fundamental principles for expression profiling of microarrays is that only DNA strands possessing complementary sequences can hybridize to each other to form a stable, double-stranded molecule. Microarrays exploit this property through the immobilization of millions of single-strand copies of a gene as individual array elements on a solid support surface. The array surface is then incubated with a mixture of labeled DNA molecules, which contains a proportional representation of all of the genes that are being expressed in a given tissue sample. Out of this mixture, only the labeled molecules that represent the same gene as the immobilized DNA elements can form heteroduplexes. By measuring the amount of label that is bound to each array element at the end of the hybridization reaction, one can determine the relative transcript abundance level of each gene. Because each microarray comprises many elements, RNA abundance levels for thousands of genes can be measured in a single experiment. By comparing abundance levels from several experiments, the investigator can then correlate patterns of gene expression with particular tissues or experimental conditions. One way of achieving this is through the identification of a statistical model of microarray hybridization intensity data that describes differential regulation, sample-sample variability, measurement noise. This data model will then be used to analyse the microarray data and develop optimal methods for detecting differentially regulated peripheral blood leukocyte mRNA from cattle infected with Trypanosoma evansi or Trypanosoma congolense using microarray in order to assay components of the immune and inflammatory responses and identify potential correlates of the pathology. This will use a 15,000 element bovine cDNA chip.

1.2 Objectives of the study

This project seeks to employ a functional genomics approach using a bovine cDNA microarray platform, to compare 'genome-wide' transcriptional responses in blood leukocytes following infection with species of Trypanosoma that differ in the severity of pathogenicity. The specific objectives are to :-

- To identify a statistical model of microarray hybridization intensity data that describes:
  a) differential regulation
  b) sample-sample variability and
  c) measurement noise.
- To use this data model to analyse the microarray data and develop optimal methods for detecting differentially regulated peripheral blood leukocyte mRNA from cattle infected with Trypanosoma congolense using microarray in order to assay components of the immune and inflammatory responses and identify potential correlates of the pathology. This will use a 15,000 element bovine cDNA chip.

1.3 Hypothesis

The best data model to use to analyse the microarray data and develop optimal methods for detecting differentially regulated peripheral blood leukocyte mRNA from infected cattle will be mixed-effects model. This is mainly because this data will be collected over time and its only mixed effects models that can overcome the violation of the assumption of error independence and also mixed effects models are very flexible in modelling the within group correlation often present in grouped data. It handles balanced and unbalanced data in a unified framework. Our hypothesis is therefore derived from the above model as follows:-

$$Y_{ijk} = \beta_{1i} + \beta_{2i} x_{ijk} + \epsilon_{ijk} \tag{1.1}$$

The above equation explains the $k-$th observation (*gene expression*) for the $j-$th-cattle in the $i-$th group. Where $i-$th group is *congolense or evansi*, expressed as 1 or 2, $j-$th cattle implies the cattle number in a particular group, expressed as 1 - 5 , and the $k-$th observation implies the different time points where the expression readings are taken, i.e point 0 - 4, expressed as 1 - 5 . The parameters β_{1i} denotes the mean for the $i-$th group which can be expressed as

$$\beta_{1i} = \beta_1 + \upsilon_{1i} \tag{1.2}$$

to represent the random intercept and

$$\beta_{2i} = \beta_2 + \upsilon_{2i} \tag{1.3}$$

to represent random slope.
Our hypothesis will therefore be derived from the two models i.e random intercept and random slope to test whether the random slopes of the graphs representing gene expressions are equal or not or if the random Intercepts of the graphs representing gene expressions are equal or not. They are formulated as follows :-
Hypothesis 1
H_0 : There is no difference in random slopes between compared groups (Boran and N'dama) H_1 : There is a difference in random slopes between compared groups (Boran and N'dama)
That is,

$$H_0 : \beta_{1i} = \beta_{2i}$$

$$against$$

$$H_1 : \beta_{1i} \neq \beta_{2i}$$

Hypothesis 2
H_0 : There is no difference in random Intercept between compared groups (Boran and N'dama)
H_1 : There is a difference in random Intercept between compared groups (Boran and N'dama)
That is,

$$H_0 : \beta_{01} = \beta_{02}$$

$$against$$

$$H_1 : \beta_{01} \neq \beta_{02}$$

Estimators of mean gene intensity in condition 0 ($Boran$) and condition 1 ($N'dama$) are derived from the equations below. (A dot in place of a subscript indicates that a the mean over this subscript has been taken)

$$Y_0 = u + t_0 + C_{(0)...} + \epsilon_{0...} \tag{1.4}$$

$$and$$

$$Y_{1...} = u + t_1 + C_{(1)...} + \epsilon_{1...} \tag{1.5}$$

which simplifies to :

$$Y_{0...} = u + t_0 \tag{1.6}$$

$$and$$

$$Y_{1...} = u + t_1 \tag{1.7}$$

because the cattle and the residual terms are assumed to have 0 mean.

1.4 Justification of the study

African trypanasomosis constrains agricultural production in areas of Africa that hold the continent's greatest potential for expanded agricultural production. Compared to animals kept in trypanasomosis free areas, animals kept in areas of moderate risk of trypanasomosis have lower calving rates, lower milk yields, higher rates of calf mortality, and require more frequent treatment with preventive and curative doses of trypanocidal drugs. At the herd level, trypanasomosis reduces milk offtake, live animal offtake and the work efficiency of oxen used for cultivation. Herds of trypanosusceptible livestock can be devastated by sudden exposure to high levels of trypanasomosis risk. In the tsetse-infested areas as a whole, trypanosomosis reduces the offtake of meat and milk by at least 50% and by generally constraining farmers from the overall benefits of livestock to farming – less efficient nutrient cycling, less access to animal traction, lower income from milk and meat sales, less access to liquid capital. Trypanosomosis costs sub-Saharan countries around US$ 10 billions annually. It is estimated that a 50% increase in the livestock population would increase the total value of agricultural production by 10%. Reid et al. (1996) estimate that about 50 million cattle are at risk of contracting tsetse-transmitted trypanosomiasis in an area of about 8.7 million km2. Winrock (1992) judge that the sub-humid zone and wetter portions of the semi-arid zone – areas in which the greatest numbers of cattle are at risk of contracting the disease - hold the continent's greatest potential for expansion of agricultural output. Current trypanosomosis control relies on trypanocidal drugs, use of trypanotolerant cattle breeds and control of the tsetse fly vector. Most heavily relied on are the trypanocidal drugs, Geerts and Holmes (1998) estimate that about 35 million doses of trypanocidal drugs are administered each year in Africa. At an average purchase price of $1 per dose, this means that African farmers are spending $35 million per year on trypanocidal drugs. The large majority of those treatments are likely given to cattle. Assuming that each animal treated was given 2 treatments per year, 17.5 million cattle were treated each year out of a total of 46 million cattle at risk. This implies that two-thirds of the cattle raised under trypanosomiasis risk were not given treatments of trypanocidal drugs. As a result, none of these methods have the full potential to work in the long-term control of the disease and this has lead to an increasing problem with resistance in the target organisms. A vaccine has been viewed as the most desirable control option. Research towards a vaccine requires the identification of immune mechanisms involved in parasite and disease control or those responses that are associated with pathogenicity. Trypanosomiasis is caused by protozoan parasites of the genus, Trypanosoma and transmitted cyclically by the tsetse fly (*Glossinaspp*). Trypanosomes are transmitted by the bite of the tsetse-fly and live in the bloodstream of infected host animals. The most important species of trypanosome that cause disease in livestock include Trypanosoma vivax, T. congolense and T. brucei brucei (Ikede, 1981), and are widely distributed in agroecological zones of Sub-Saharan Africa. Other species, T. simiae and T. evansi cause diseases in pigs (Leach and Roberts, 1965) and camels (Scot, 1973) respectively. DNA microarrays, first introduced in mid 1990's, measures the activity level of each gene (*its expression level*), in a particular cell at a particular time by measuring the relative amount of mRNA expressed by each gene. They are now mainstays of drug discovery research. Microarrays are also beginning to revolutionize

how scientists explore the operation of normal cells in the body and the molecular aberrations that underlie medical disorders. These tools promise to pave the way for faster, more accurate diagnoses of many conditions and to help scientists personalize medical care-that is, tailor therapies to the exact form of disease in each person or animal and select the drugs likely to work best, with the mildest side effects, in those individuals. The arrays come in several varieties, but most measure gene expression in a tissue sample, and consist of a lawn of double-stranded DNA molecules (*probes*) that are tethered to a glass substrate no bigger than a microscope slide. Microarrays capitalize on a convinient property of DNA: complementary base pairing. DNA is the universal genetic (*apart form RNA viruses*) material that contains the code that constitutes the blueprints for proteins. It consists of four building blocks, usually referred to by the first letter of their distinguishing chemical bases: A, C, G and T. The base A in double strand of DNA will pair only with T (*A's complement*) on another strand, and C will pair only with G. This process of pairing is known as hybridization. When such pairing occurs in a microarray it reflects a particular gene was active or expressed in the sample. The images produced by a scanner (*a machine which measures fluorescence – the amount of hybridization for each gene probe*) from the pairing, normally come in four colours which represent the following:

1. Red Colour :Gene that strongly increased activity in infected cells (Up-regulation)
2. Green Colour :Gene that strongly decreased activity in infected cells (down-regulation)
3. Yellow Colour :Gene that was equally active in infected cells and uninfected cells
4. Black Colour :Gene that was inactive

These images produced by microarray hybridization and scanning, normally form a complex and a large amount of data. Due to the complex nature and sheer amount of data produced from microarray experiments, standardized systems and tools for data management are needed in order to publish the results in a proper and sound way as well as to be able to benefit from other publicly available gene expression data. As a result, data analysis is needed to extract meaningful information from the digital images thus facilitating automatic understanding of the contents of the images. Since the biological and technical intricacies of microrarray experiments are not easily accessible to analysis , these analyses seek to organize results in a meaningful order i.e flags, controls and experiments are pointed out and checked. Variation due to systematic errors is removed and data from different slides are compared. Statistics is needed in many steps during analysis. Statistical tools will be used for evaluation of the raw data during the above-mentioned steps and in further analysis to find significantly differentially expressed genes. In these further analysis steps, statistically significant , quality-checked data will be separated from not interesting and not-trustworthy data. The next step will be to find differentially expressed genes using statistical tools or to group the good quality data (*usually only a small fraction of the original raw data*) into meaningful clusters by e.g clustering. The goal of clustering will mainly be to find similarly behaving genes or patterns related to time scale, time point, developmental phase of the disease or the different treatments of the cDNA sample. Finally, we will have to link the observations to biological data, regulation of genes and to annotations of functions and biological processes in order to

1. Identify genes whose expression changed with time.
2. Be able to correlate gene expression to parasitology and immune cell types.
3. Identify differences between infecting patterns.
4. Identify genes that changed during course of infection and find out whether they are the same.
5. Ascertain what value the gene expression have to the making of a vaccine.

1.5 Literature review

Mixed-effects models are primarily used to describe relationship between a response variable and some covariates in data that are grouped according to one or more classification factors. Such grouped data include longitudinal data (), repeated measures data (), multilevel data () and block designs ().
Pinheiro and Bates (2000) introduced basic concepts of linear mixed effects models through the analysis of several real data examples. In their work, a data by Devore (2000, $example10.10pg427$) was used as an example. This is an experiment in non-destructive testing for longitudinal stress in railways rails from an article in Materials Evaluation.
In this experiment, six rails were chosen at random and tested three times each by measuring the time it took for a certain type of ultrasonic wave to travel the length of the rail. This was their example of a one-way classification since they had only one experimental setting (*the rails*) that changed between the observation. They fitted this data as a fixed effect model and as a random effect model. Out of their working, they found that the fixed effects model accounted only for rail effects but could not provide a useful representation of the rails data.
In other words, fixedeffects model did not provide an estimate of the between rail variability thus could not provide information on the population of rails from which the sample was taken. This problem was overcomed by using a random effects model by treating the rail effects as random variations around a population mean. i.e

$$Y_{ij} = \beta + (\beta_i - \beta) + \epsilon_{ij} \tag{1.8}$$

where $\beta = \Sigma\frac{\beta_i}{6}$ represents the average travel time for the rails in the experiment.
Equation (1.8) reduces to:

$$Y_{ij} = \beta + b_i + \epsilon_{ij} \tag{1.9}$$

where:

1. β is the mean travel across the population of rails being sampled,

2. b_i is a random variable representing the deviation from the population mean of the mean travel time for the ith rail,

3. ϵ_{ij} is a random variable representing the deviation in travel time for observation j on rail I from the mean travel time for the ith rail.

In their fit and analysis to estimate the parameters, they used the S and S-plus codes.
For a two classification factors, Pinheiro and Bates (2000) took a data from ergometric experiment that had a randomized block design. The two classification factors are *experimental factor* (for which we use fixed effects) and *blocking factor* (for which we use random effects). In this experiment, the experimenters recorded the effort required by each of the nine different subjects to arise from each of the four types of stools. In their analysis to compare the four particular types of stools, they used fixed effects for the **Type factor** and for the nine different subjects representing a sample from the population about which they would make their inferences, they used random effects to model the **Subject factor.**
A model with fixed effetcs β_j for the **Type factor** and random effects b_i for the **Subject factor** could be written as :

$$Y_{ij} = \beta_j + b_i + \epsilon_{ij},\ i = 1, \ldots, 9,\ j = 1, \ldots, 4, \tag{1.10}$$

where $b_i \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$.
To fit and estimate the parameters of this unreplicated mixed-effect model, they used the S-Plus software.
To account for the replicate measurements, Pinheiro and Bates (2000) used the Machine data in Milliken and Johnson (1992) which gives the productivity score for each of six randomly chosen workers tested on each of three different machine types. Each worker used each machine three times giving three replicates at each set of conditions. This gave strong indications of differences between machines and some indications of differences between workers, but very little variability in the productivity score for the same worker using the same machine.
Replications allowed them to asses presence of interactions between worker and machine. They modeled the *subject or worker factor* with random effects and the *type or machine factor* with fixed effects.
An interaction plot in S was drawn. The model incorporating the random interaction terms, $b_{ij}, i = 1, \ldots, 6, j = 1, \ldots, 3$ is

$$Y_{ijk} = \beta_j + b_i + b_{ij} + \epsilon_{ijk}, i = 1, \ldots, 6, j = 1, \ldots, 3 \text{ and } k = 1, \ldots, 3, \tag{1.11}$$

where
$b_i \sim N(0, \sigma_1^2)$,
$b_{ij} \sim N(0, \sigma_2^2)$, and
$\epsilon_{ijk} \sim N(0, \sigma^2)$
in which case there are random effects at two levels:

- the effects b_i for the work and

- the effects b_{ij} for the type of machine within each worker

In their fit,to estimate the parrameters of this model,by the MLE and REML criterions,they used the S and S-Plus softwares. The Machines data above are balanced i.e every worker is tested on every machine exactly three times.
Millikens and Johnson,(1992) analysed this data both as balanced and unbalanced. They randomly deleted ten observations from the balanced data above. To analyse the estimates by the MLE, they used SAS PROC Software.
Pinheiro and Bates, (2000) Used S and S-Plus to produce sensible MLE or REML estimates from the unbalanced data.
For random effects analysis of covariance models, Pinheiro and Bates, (2000), used a data given in Potthoff and Roy,(1964) where a set of measurements of the distance from the pituitary gland to the pterygomaxillary fissue taken every two years from 8 years of age to 14 years of age on a sample of 27 children (16 *males and* 11 $females$). Taking the data for the females only, they fitted separate linear regression models for each girl and examined individual confidence intervals on the parameters using S and S-Plus softwares. They used the same data to extract the best linear unbiased predictions (BLUPs) of the random effects from the fitted model using the same software.
Pinheiro and Bates, (2000) also analysed experiments in models for nested classification factors and split plot.
In nested classification factors, they analysed a data from an experiment on the pixel intensity in computerised tomography (CT) scans in which experimenters injected each of the 10 dogs with a dye contrast then recorded the mean pixel intensities from CT scans of the right and left lymph nodes in the axillary region on several occasions up to 21 days post injection. In this experiment, the left and right sides were expected to be diffenrent but the difference was not expected to be systematic in terms of left and right;i.e, for one dog the left side may have greater pixel intenities than the right, while for another dog the opposite may be true. The dog and side were considered to be nested classification factors. For the analysis of this experiment, they used S and S-Plus software.
More work on scientific experiments (*medical experiments*) was analysed by Geert Verbeke and Geert Molenberghs, (2000). One of their experiments, based reffered to as the rat data was setup at the Department of Orthodontics of the Catholic University of Leuven (KUL) in Belgium. To investigate the effect of inhibition of the production of testosterone in male $Wistar\ rats$ on their cranofacial growth. 50 male $Wistar\ rats$ had been rrandomised to etiher a control or one of the two treatment groups where treatment consisted of a low or high dose of the drug Decapeptyl, which is na inhibitor for testosterone production in rats.
The responses of interests are distances (*in pixels*) between well defined points on X-ray pictures of the scan of each rat, taken after the rat had been anaesthetised. Of major interest wa sthe estimation of changes over time and testing whether these changes were treatment dependent. In their work, they considered one of the measurements which could be used to charcterise the height of the skull. For the anlysis of this experiment, a linear mixed effects model was fitted and SAS program in was used in estimation of the parameters.

Another experiment analysed by Geert Verbeke and Geert Molenberghs, (2000), was the toenail data (TDO). This was data obtained from a randomised,double-blind, parallel-group, multi-centre study for the comparison of two oral treatments for *toenail dermatophyte onychomycosis* ($T1$, described in full detail by De Backer *etal*, (1996). Tdo is a common toenail infection, difficult to treat, affecting more than 2 percent (Roberts,(1992)). To analyse all their experiments, SAS program was used to estimate all the parameters.
A randomized complete block experiment, with six blocks of three main plots of equal size was carried out in this research station.Maize was planted using densities of two, three or four seeds per hole, and planting density was randomised to the plots such that there was one plot of each planting density per block. Spacing between holes was the same for all plots. Thinning was carried out on plots planted with more than two seeds per hole - at 8 and 14 weeks for the plots with 4 seeds per hole and at 14 weeks for the plots with 3 seeds per hole - so that from 14 weeks onwards all plots had two seeds per hole. Two thinning practices were used:

1. Removal of the smallest plant from the hole.

2. Removal of the second largest plant from the hole.

To incoporate this into the experiment, the 3-seeds and 4-seeds plots were subdivided into two and these two subplots were randomly allocated to one of the different thinning practises.The 2 seed plot was not subdivided. At the 8^{th} *and* 14^{th} weeks the amount of green forage ($Kg/ha\ of\ dry\ matter$) was recorded for the plots which were thinned. At the end of the study,at the 28^{th} week, green yield ($KgDm/ha$) was determined for all plots.
Their second experiment was on concentrate feeding trial which was to test the feasibility of changing farmers' concentrate allocation practice by shifting the concentrates to early lactation. The data comprised weekly records of milk yields and concentrates offered for the first 12 weeks of lactation, and approximately fortnightly thereafter. Data were collected between March 1999 and March 2000,and complete datasets were achieved for 65 cows belonging to 53 households and calving between March and September 1999. The objective of the analysis was to determine:

(a) The influence of house hold ($farm$) and cow factors on milk yield and

(b) Their relationships between milk yield and concentrate fed at different phases of lactation.

Five six-week sampling periods upto 30 weeks of lactation were defined, and average daily milk yield (kg/day) in these six-week intervals was estimated for each animal from fitted lactation curves.
Their last example was a sheep breeding trial where data used was from the kenyan coast between 1991 and 1997.(Baker,1998). They were to compare the genetic resistance to Helminthiasis of two indegenous breeds of sheep - Dorper (D) and Red Maasai(R)- and to use this information alongside survival rate to compare the overall productive performance of each breed. Throughout the six years, Dorper (D), Red Maasai(R) and R x D ewes were mated to Red Maasai and Dorper rams to produce a number of different lambs genotypes. Only four offspring genotypes were considered:D X D, D X R, R X D, R X R.

In all these three examples,Allan and Rowlands,(2001) made use of the Genstat output for the mixed models in their analysis.
African bovine trypanasomosis, caused by the protozoan parasite Trypanosoma congolense, is endemic throughout sub-Saharan Africa and is a major constraint on livestock production. A promising approach to disease control is to understand and exploit naturally evolved trypanotolerance.We describe the first attempt to investigate the transcriptional response of susceptible Boran (*Bos indicus*) cattle to trypanosome infection via a functional genomics approach using a bovine total leukocyte (*BOTL*) cDNA microarray platform. Four male Boran cattle were experimentally infected with T. congolense and peripheral blood mononuclear cells (*PBMC*) were collected before infection and 13, 17, 23 and 30 days post-infection (*dpi*). A reference experimental design was employed using a universal bovine reference RNA pool. Data were normalised to the median of a set of invariant genes (*GAPDH*) and BRB-Array tools was used to search for statistically significant differentially expressed genes between each time-point. Using a set of 20 microarray hybridisations, we have made a significant contribution to understand the temporal transcriptional response of bovine PBMC in vivo to a controlled trypanosome infection. The greatest changes were evident 13 dpi after parasites were first detected in the blood. (Emmeline W. Hill,*et al* (2003)) In Africa, trypanasomosis is a tsetse-transmitted disease which represents the most important constraint to livestock production. Several indigenous West African taurine (*Bostaurus*) breeds, such as the Longhorn (*N'Dama*) cattle are well known to control trypanosome infections. This gene-based ability called 'trypanotolerance' results from various biological mechanisms under multigenic control. The methodologies used so far have not succeeded in identifying the complete pool of genes involved in trypanotolerance. New postgenomic biotechnologies such as transcriptome analyses are efficient in characterizing the pool of genes involved in the expression of specific biological functions. We used the serial analysis of gene expression (*SAGE*) technique to construct, from peripheral blood mononuclear cells of an N'Dama cow, two total mRNA transcript libraries, at day 0 of a Trypanosoma congolense experimental infection and at day 10 post-infection, corresponding to the peak of parasitaemia. Bioinformatic comparisons in the bovine genomic databases allowed the identification of 187 up- and down-regulated genes, EST and unknown functional genes. Identification of the genes involved in trypanotolerance will allow the setting up of specific microarray sets for further metabolic and pharmacological studies and the design of field marker-assisted selection by introgression programmes. (Berthier,*et al*(2003)) Parasitic diseases caused by protozoan and helminth parasites are among the leading causes of morbidity and mortality in tropical and subtropical regions of the world. Unfortunately, at present, there is no vaccine against any human parasitic disease. Conventional vaccine methods have largely failed against parasitic infections. This is due, in part, to the complexity of the parasite life cycle, the ability of the parasite to evade the immune system, and difficulties in identifying and eliciting the desired protective immune responses. The discovery of DNA vaccines has renewed hope for vaccine development against parasites. In the last decade, DNA vaccines were successful in inducing at least partial protection against several parasitic diseases. This review discusses the latest developments in DNA vaccines against tropical parasitic diseases.(*Akram A Da'dara and Donald A Harn*) There is consensus that vaccination will provide the most effective means for the control of trypanosomosis, but no vaccine

is available. In addition, sensitive molecular tools are also required for fine-scale identification and more detailed characterisation of all T. brucei subspecies, and for epidemiological investigations of African trypanosomosis. Such specific diagnostic tool will also provide necessary adjunct to the effective use of any new drug or vaccine products. Recent developments towards developing specific markers through fine-scale trypanosome genome analysis (Agbo *et al.*, 2001; Agbo *et al.*, 2002; Agbo *et al.*, In press) indicate that specific molecular diagnosis of T. brucei subspecies is feasible. There are high expectations that genomics will offer new leads to rationale vaccine design and accelerate the discovery of novel drug targets, as well as the development of practical and robust diagnostic assays .

1.6 Data and methodology

1.6.1 Data

20 Boran and 20 N'Dama cattle were infected with Trypanosoma congolense IL1180. Challenge was by bite of 8 infected Tsetse flies. Animals were in 4 groups labelled T1, T2, T3 and T4 each consisting of 5 Boran and 5 N'Dama in pairs. Each pair being made of two animals that were age and sex matched and had similar preinfection clinical profiles. Liver biopsies were taken from each animal at least two weeks before infection. After infection, liver, spleen, lymph node and other tissues specimen were collected by biopsy or at postmortem as shown in figure 1. So for each animal there is a preinfection liver sample and up to two postinfection biopsy samples. Cattle in Group T4 were sacrificed after 21 days and Cattle in Group T1 were sacrificed 35 day after infection. During the Infection, Clinical data including body weight, rectal temprature, packed cell volume (PCV), and parasitemia were determined 3 times each week.

a) Clinical parameters including temperature [daily from day 0], and PCV [from day 0, three times each week]. This is to help measure disease with respect to pack cell volume. i.e the lower the PCV, the higher the disease.

b) Differential blood cell count through microscopy slides [daily from day 0]

c) FACS analysis of buffy coat leukocyte populations [from day 0, every four days]

d) Serology (acute phase protein ELISA) using ELISA vs. Tryps Antigen [from day 0, three times each week]

e) Trypanosomes by buffy coat microscopy and blood count (parasitaemia count / ml) [daily from day 3]

f) DNA content of trypanosomes by buffy coat FACS analysis [from day 7, three times each week].

1.6.2 Methodology

Image processing and computational analysis

The microarray image produced by the scanner will be the raw data of the project. The first step will be to convert the digital TIFF images of the hybridisation intensity generated by the scanner into numerical measures of the hybridisation intensity of each channel on each feature. Measures to be used to help in calculating the intensity for each feature are:

- Signal mean
- Background mean
- Signal median
- Background median
- Signal standard deviation
- Background standard deviation
- Diameter
- Number of pixels
- Flag

Normalisation

This is a term for a collection of methods that are directed at resolving the systematic errors and bias introduced by the microarray experimental platform. The microarray data that will be generated by the feature extraction software (GenePix Pro) will typically be in the form of one or more text files.

Analysis

The purpose of microarray experiments is to measure and compare gene expression profiles in different types of tissue. A variety of discriminative data analysis methods have been exploited to identify groups of genes with correlated expression profiles across experimental conditions or to reveal specific patterns of gene expression for phenotype classification and prediction and potentially for medical diagnostics.
These methods are specifically designed to take full advantage of the highly parallel nature of microarray data. They measure the correlation of expression profiles of thousands of genes simultaneously. These methods can be very insightful as they are designed to reveal complex pattern

of gene expression. In this analysis, I focus on a simpler problem of individually identifying differentially regulated genes. Such an approach can provide more detailed information about single gene intensity observations.
Methods like exploratory data analysis will be used for gaining insight into the data.
EDA employs graphical and quantitative techniques. These are sequential plots,histograms,box-plots among others. This normally gives a better insight to the dataset allowing it to reveal its underlying structure,detect outliers and anomalies.
Statistics for testing goodness-of-fit like -2 Log likelihood,Akaike's Information Criterion and Bayesian Information Criterion will be used.

1.7 Expected output

At the end of this work we expect to be able

- To identify a statistical model of microarray hybridization intensity data that describes:
  a) differential regulation
  b) sample-sample variability and
  c) measurement noise.
- To use this data model to analyse the microarray data and develop optimal methods for detecting differentially regulated peripheral blood leukocyte mRNA from cattle infected with Trypanosoma congolense using microarray.
- We also expect to see the expressions of Boran are Higher than the Nd'ama since Boran are susceptible to tryps and Nda'ma are resistant.

Chapter 2

Literature review

2.1 The hierarchical linear Model

This section presents the general hierarchical linear model which allows intercepts as well as slopes to vary randomly. It follows the approach of a two-level nesting structure and the level-one units are called - for convenience - 'individuals', while the level-two units are called 'groups'. The notation is also the same.

2.1.1 Random slopes

In the random intercept model of our data,the groups differ with respect to the average value of the dependent variable:the only random group effect is the random intercept. But the relation between explanatory and dependent variables can differ between groups in more ways. In the analysis of covariance,this phenomenon is known as heterogeneity of regressions across groups, or as group-by-covariate interaction. In the hierarchical linear model, it is modeled by $random-slopes$.
By considering the model with group-specific regressions of Y on one level-one variable X only, but without the effect of Z below;

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + R_{ij} \tag{2.1}$$

The intercepts β_{0j} as well as the regression coefficients, or slopes, β_{1j} are group-dependent. These group-dependent coefficients can be split into an average coefficient and the group-dependent deviation:

$$\beta_{0j} = \gamma_{00} + U_{0j} \tag{2.2}$$

$$\beta_{1j} = \gamma_{10} + U_{1j} \tag{2.3}$$

Substitution leads to the model

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + U_{0j} + U_{1j}x_{ij} + R_{ij} \tag{2.4}$$

It is assumed here that the level-two residuals U_{0j} and U_{1j} as well as the level-one residuals R_{ij} have means 0, given the values of the explanatory variable X. Thus, γ_{10} is the average regression coefficient just like γ_{00} is the average intercept. The first part of (2.4), $\gamma_{00} + \gamma_{10}x_{ij}$, is called the *fixed-part* of the model and the second part, $U_{0j} + U_{1j}x_{ij} + R_{ij}$, is called the *randompart*.
The term $U_{1j}x_{ij}$ can be regarded as a *random interaction between group and X*. This model implies that the groups are characterized by two random effects: their intercept and their slope. We say that X has a random slope, or a random effect,or a random coefficient. These two group effects will usually not be independent, but correlated. It is assumed that, for differet groups, the pairs of random effects (U_{0j}, U_{1j}) are independent and identically distributed, that they are independent of the level-one residuals R_{ij}, and that all R_{ij} are independent and identically distributed. The variance of the level-one residuals R_{ij} is again denoted σ^2; the variances and covariance of the level-two residuals (U_{0j}, U_{1j}) are denoted as follows:

$$\begin{array}{lcl} var(U_{0j}) & = & \tau_{00} = \tau_0^2; \\ var(U_{1j}) & = & \tau_{11} = \tau_1^2; \\ var(U_{0j}, U_{1j}) & = & \tau_{01}. \end{array} \tag{2.5}$$

Therefore one can say that the unexplained group effects are assumed to be exchangeable.

2.1.2 Explanation of random intercepts and slopes

Regression analysis aims at explaining variability in the outcome (i.e dependent) variable. Explanation is expalined here as being able to predict the value of the dependent variable from knowledge of the values of the explanatory variables. The unexplained variability in single-level multiple regression analysis is just the variance of the residual term. Variability in multilevel data, however, has a more complicated structure. Explaining variability in a multilevel structure can be achieved by explaining variability between individuals but also by explaining variability between groups;if there are random slopes as well as random intercepts,at the group level we could try to explain the variability of slopes as well as intercepts.
In the model defined by (2.1)-(2.4), some variability in Y is explained by the regression on X, i.e,by the term $\gamma_{10}x_{ij}$; the random coefficients U_{0j}, U_{1j} and R_{ij} each express different parts of the

unexplained variability. In the first place, we will try to find explanations in the population of individuals (at level one). The part of residual variance that is expressed by $\sigma^2 = var(R_{ij})$ can bediminished by including other level-one variables. Since group compositions with respect to level-one variables can differ from group to group,inclusion of such variables may also diminish residual variance at the group level. A second possibility is to try to find explanations in the population of groups (at level two). If we wish to reduce the unexplained variability associated with U_{0j} and U_{1j}, we can also say that we wish to expand equations (2.2) by predicting the group-dependent regression coefficients β_{0j} and β_{1j} from level-two variables Z. Supposing we have one such variable, this leads to regression formulae for β_{0j} and β_{1j} on the variable Z,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + U_{0j} \tag{2.6}$$

$$\beta_{0j} = \gamma_{10} + \gamma_{11} Z_j + U_{1j} \tag{2.7}$$

Chapter 3

Data and methodology

3.1 Data

20 Boran and 20 N'Dama cattle were infected with Trypanosoma congolense IL1180. Challenge was by bite of 8 infected Tsetse flies. Animals were in 4 groups labelled T1, T2, T3 and T4 each consisting of 5 Boran and 5 N'Dama in pairs. Each pair being made of two animals that were age and sex matched and had similar preinfection clinical profiles. Liver biopsies were taken from each animal at least two weeks before infection. After infection, liver, spleen, lymph node and other tissues specimen were collected by biopsy or at postmortem as shown in figure 1. So for each animal there is a preinfection liver sample and up to two postinfection biopsy samples. Cattle in Group T4 were sacrificed after 21 days and Cattle in Group T1 were sacrificed 35 day after infection. During the Infection, Clinical data including body weight, rectal temprature, packed cell volume (PCV), and parasitemia were determined 3 times each week.

a) Clinical parameters including temperature [daily from day 0], and PCV [from day 0, three times each week]. This is to help measure disease with respect to pack cell volume. i.e the lower the PCV, the higher the disease.

b) Differential blood cell count through microscopy slides [daily from day 0]

c) FACS analysis of buffy coat leukocyte populations [from day 0, every four days]

d) Serology (acute phase protein ELISA) using ELISA vs. Tryps Antigen [from day 0, three times each week]

e) Trypanosomes by buffy coat microscopy and blood count (parasitaemia count / ml) [daily from day 3]

f) DNA content of trypanosomes by buffy coat FACS analysis [from day 7, three times each week].

Microarray data will be collected through a series of laboratory steps and by using image acquisition software before its ready for analysis. The steps are as follows:

1. Sample preparation and Labelling

   The first step is to extract RNA from the tissue of interest(purified buffy coat leukocytes). There is much variability among the individual scientist performing the extraction. The DNA will be fluorescent labelled with two dyes Cy3(excited by a green laser) and Cy5 (excited by a red laser). The two samples will be hybridised to the array one labelled with each dye, this will allow the simultaneous measurement of both samples.

2. Hybridisation

   Is the step in which the DNA probes on the glass and the labelled RNA target form heteroduplexes.

3. Washing

   After hybridisation, the slides are washed . The reason for this is to remove excess hybridisation solution from the array. This ensures that the only labelled target on the array is the target that has specifically bound to the features on the array and thus represents the DNA that we will be measuring. The second reason is to increase the stringency of the experiment by reducing cross-hybridisation. This will be achieved by washing in a low-salt wash or with a high-temperature wash.

4. Image Acquisition

   The final step of the lab process will be to produce images of the surface of the hybridised array. The heteroduplexes on the array where the target will have bound to the probe will contain dye that fluoresces when excited by light of an appropriate wavelength. The slide will be placed in a scanner which is a device that reads the surface of the slide. The scanner to be used will be the GenePix Microarray Scanner. It's mainly preferred since one can acquire and analyse images of arrays without having to learn the finer details of image processing. Each pixel on the digital image represents the intensity of fluorescence induced by focussing the laser at that point on the array. The dye at that point will be excited by the laser and will fluoresce; this fluorescence is detected by a photomultiplier tube (PMT) in the scanner. In order to scan the whole array, the laser must be focussed on every point on the array. This will be achieved either by moving the slide so that the laser can focus on different points or by shifting the optics to achieve the same result. The colour images are usually stored at tagged image file format ($TIFF$).The array data is stored in 16 bits. This means that the intensity of each pixel in each channel is quantified as a 16-bit number, which takes values between 0 and 216 - 1, which is equal 65,535.Since background is approximately 100, and saturation can occur when the average pixel intensity

is larger than 50,000, the microarray can detect intensities over an approximately 500- fold dynamic range. With 10-m pixel sizes, a typical microarray image will be 7,500 x 2,200 pixels. This means that each of the two TIFF images is 32 Mb, these are large files so in order to produce a large number of microarray images data storage becomes an important consideration. Pixel resolution of the image will be chosen so that each feature has sufficient pixels to make the measurement of the intensity of the feature robust from pixel to pixel noise. The recommended pixels per feature are at least 50 pixels.

3.1.1 Variables

Time-fixed factors

The contribution of a set of fixed factors (not time-varying) to gene expression change will be considered. These factors included the breed group and and the tissue of the particular animal within a specific breed group.

Time-varying factors

These represent exposure in the interval before each measurement of gene expression. The use of time point variable (day) attempts to account for temporality assumptions. For each day, we will calculate the mean number of gene expression score.

3.2 Methodology

3.2.1 Image processing

The microarray image produced by the scanner will be the raw data of the project. The first step will be to convert the digital TIFF images of the hybridisation intensity generated by the scanner into numerical measures of the hybridisation intensity of each channel on each feature. Four steps will be involved and these are:

1. Identifying the positions of the features on the microarray
2. For each feature, the pixels on the image that are part of the feature will also be identified.
3. For each feature, nearby pixels that will be used for background calculation will be identified.

4. Perform calculation of the numerical information for the intensity of the feature, the intensity of the background and quality control information.

Below are the measures which will help in calculating the intensity for each feature.

- Signal mean: the mean of the pixels comprising the feature. It will be used for measuring hybridisation intensity

- Background mean: the mean of the pixels comprising the background around the feature. It will be subtracted from the feature intensity to provide a reliable estimate of hybridisation intensity to each feature

- Signal median: the median of the pixels comprising the feature. It will be used for measuring hybridisation intensity its preferable to use over signal mean since its more robust to outlier pixels than the mean. A small number of very bright pixels (arising from the noise) will have the potential to skew the mean, but will leave the median unchanged.

- Background median: the median of the pixels comprising the background. It will be subtracted from the feature intensity to provide a more reliable estimate of hybridisation intensity to each feature

- Signal standard deviation: the standard deviation of the pixels comprising the feature. It will be used as a quality control for the array. I.e. if the standard deviation of a feature is greater than say 50% of the median intensity, the feature could be rejected as substandard.

- Background standard deviation: the standard deviation of the pixels comprising the background.

- Diameter: the number of pixels across the width of the feature.

- Number of pixels: number of pixels comprising the feature

- Flag: a variable that is 0 if the feature is good and will take different values if the feature is not good.

Normalisation

This is a term for a collection of methods that are directed at resolving the systematic errors and bias introduced by the microarray experimental platform. The microarray data that will be generated by the feature extraction software (GenePix Pro) will typically be in the form of one or more text files. Before we use this data to answer our scientific questions, the following steps have to be done;

- Removing flagged features : These are features which GenePix Pro will have detected some type of problem in them and will need to be removed.

- Background subtraction: This involves subtracting the background signal from the feature intensity. This background signal is thought to represent the contribution of non-specific hybridisation of labelled target to the glass as well as the natural fluorescence of the glass slide itself.

3.2.2 computational analysis

The purpose of microarray experiments is to measure and compare gene expression profiles in different types of tissue. A variety of discriminative data analysis methods have been exploited to identify groups of genes with correlated expression profiles across experimental conditions or to reveal specific patterns of gene expression for phenotype classification and prediction and potentially for medical diagnostics.
These methods are specifically designed to take full advantage of the highly parallel nature of microarray data. They measure the correlation of expression profiles of thousands of genes simultaneously. These methods can be very insightful as they are designed to reveal complex pattern of gene expression. In this analysis, I focus on a simpler problem of individually identifying differentially regulated genes. Such an approach can provide more detailed information about single gene intensity observations.
The critical point in methods addressing questions about single gene observations is assessing the variability of intensity observations. However, its only in experimental designs with repeated measurements that is possible to calculate an accurate estimate of the observation variability for each gene.

Exploratory Data Analysis

Before fitting a linear model, its always very important to confirm whether the data conforms to the underlying assumptions of a linear model. This is best achieved through exploratory data analysis (EDA).
EDA employs graphical and quantitative techniques.These are sequential plots,histograms,box-plots among others.This normally gives a better insight to the dataset allowing it to reveal its underlying structure,detect outliers and anomalies.

Statistics for testing goodness-of-fit

-2 Log likelihood

A measure of how well the model fits the data, also called the deviance. The smaller the value the better the fit.

Finite sample corrected AIC (AICc)

$$\mathbf{A}ICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{3.1}$$

where n is the number of observations (i.e. the sample size) and k isthe number of repressors, including the constant.
AICc is a measure for selecting and comparing mixed models based on the -2 (Restricted) log likelihood. Smaller values indicate better models. The AICc "corrects" the AIC for small sample sizes. As the sample sizes. As the sample size increases, the AICC converges to the AIC. AICc should be employed regardless of sample size (Burnham and Anderson,2004)

Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) is a statistical criterion for model selection. It's a measure for selecting and comparing mixed models based on the -2 (Restricted) log likelihood. Smaller values indicate better models. The BIC also penalizes overparametrized models, but more strictl than the AIC.
The formula for the BIC is given below. This assumes that the model errors are normally distributed.

$$\mathbf{B}IC = ln\frac{RSS}{n} + k\frac{lnn}{n}, \tag{3.2}$$

where n is the number of observations (i.e. the sample size), k is the number of repressors, including the constant, RSS is the residual sum of squares from the estimated model and L is the maximized value of the likelihood function for the estimated model.

Akaike's Information Criterion (AIC)

Akaike's information criterion (AIC) developed by Hirotsugu Akaike in 1971 and proposed in Akaike (1974), is a measure of the goodness of fit of an estimated statistical model. The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data. It is a measure for selecting and comparing mixed models based on the -2 (Restricted) log likelihood. Smaller values indicate better models.
In the general case,the formula for AIC is given as

$$\text{AIC} = 2k - 2ln(L)$$

where k is the number of parameters and L is the likelihood function.

Assuming that the errors are normally and independently distributed, the AIC formula becomes

$$AIC = 2k + n \log L \frac{RSS}{n}$$

where n is the number of observations and RSS is the residual sum of squares.
Increasing the number of free parameters to be estimated improves the goodness of fit, regardless of the number of free parameters in the data generating process. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting. The preffered model will be the one with the lowest AIC value.
A linear model is a model in which the expected value of a random variable is expressed as a linear function of the parameters of the model. I am particularly interested in linear models because;

a) This class of model has been extensively used and studied for more than 70 years.

b) Linear models are generally easy to interpret since each term may represent some physical or physiological variable.

c) They are flexible and easy to modify to serve a particular purpose.

d) Applications of linear models to the analysis of microarray data have been published.

This entire output of a microarray experiments is represented in a single equation with one single set of parameters. In particular, differences in intensities across genes are accounted for by a term in the model. These global models can provide very interesting insights into the results of microarray experiments.
However, they do not permit a fine estimation of the variability of gene intensity observation they try to encompass all the data in one single equation. This is mainly because two major sources of variability exist in microarray experiments. The first cause for variation is physiological: genes may be expressed differently in different tissue samples even if they are nominally in the same condition regarding the biological question of interest. This source of variation is often reffered to as "biological" variance. The second source of fluctuation is the uncertainty inherent to the microarray technology and experimental protocol. Below is a schema illustrating how microarray data is broken down for analysis with the linear model.
Linear models accounting for more than one source of uncertainty in the data are called mixed-effect linear model. I propose a mixed-effect linear model based on explicit estimation of a regulation term i.e. difference between two conditions, (tissue sample to tissue sample) variability term and a residual measurement error term for each gene. This is because we can't use ANOVA since the gene expression data which will be measured over time is mainly from a particular animal and the assumption that the observations will be independent of each other wont be mate since the subsequent observation are dependent of the previous ones.
The mixed effect linear model for gene intensity observations is developed for experimental designs with several observations in each condition, the repeated measurements involved to produce

the longitudinal data. The model will be used to test hypotheses about the statistical significance of differences in gene intensity /expression between Boran and N'dama tissues, and simulate data sets under well-controlled conditions.
In practise,longitudinal data are often higly unbalanced in the sense that not an equal number of measurements is available for all subjects and / or that measurements are not taken at fixed time points.Due to their unbalanced nature,many longitudinal data sets cannot be analyzed using multivariate regression techniques. A natural alternative arises from observing that subject-specific longitudinal profiles can often be well approximated by linear regression functions. One hereby summarizes the vector of repeated measurements for each subject by a vector of a relatively small number of estimated subject-specific regression coeffiecients. Afterward, in a second stage, multivariate regression techniques can be used to relate these estimates to known covariates such as treatment, disease classification, baseline characteristics, and so forth.The general linear mixed model is introduced as a result of combining the two stages into one single statistical model.

Two-Stage Analysis (Stage 1)

Let the random variable Y_{ij} denote the (possibly transformed) gene expresion, for the $i-$th cattle, measured at time $t_{ij}, i = 1, \ldots, N,\ j = 1, \ldots,\ n_i$ and let $\mathbf{Y}_i$ be the n_i-dimensional vector of all repeated measurements for the $i-$th subject,that is $\mathbf{Y}_i = Y_{i1},\ Y_{i2},\ \ldots,\ Y_{ini}$. The first stage of the two-stage approach assumes that $\mathbf{Y}_i$ satisfies the linear regression model

$$\mathbf{Y}_i = Z_i\beta_i + \epsilon_i \tag{3.3}$$

where Z_i is a $(n_i\ x\ q)$ matrix of known covariates, modeling how the response evolves over time for the $i-$th subject.Further,β_i is a q-dimensional vector of unknown subject-specific regression coefficients, and ϵ_i is a vector of residual components $\epsilon_{ij},\ j = 1, \ldots, n_i$. It is usually assumed that all ϵ_i are independent and normally distributed with mean vector zero, and covariance matrix $\sigma^2\mathbf{I}_{ni}$, where $\mathbf{I}_{n_i}$ is the n_i-dimensional identity matrix.

Stage 2

In a second step,a multivariate regression model of the form

$$\beta_i = K_i\beta + \mathbf{b}_i, \tag{3.4}$$

is used to explain the observed variability between the subjects, with respect to their subject-specific regression coefficients β_i. K_i is a (q x p) matrix of known covariates,and β is a p-dimensional vector of unknown regression parameters. Finally, the $\mathbf{b}_i$ are assumed to be independent,following a q-dimensional normal distribution with mean vector zero and general covariance matrix D.

The Model

In order to combine the models from the two-stage analysis, we replace β_i in (3.3) by (3.4),yielding

$$\mathbf{Y}_i = X_i\beta + Z_i b_i + \epsilon_i \tag{3.5}$$

Where $\mathbf{X}_i = Z_i K_i$ is the appropriate $(n_i \times p)$ matrix of known covariates, and where all other components are as defined earlier. Model given by (3.5), above is called a linear mixed (-effects) model with fixed effects β and with subject-specific effects $\mathbf{b}_i$. It assumes that the vector of repeated measurements on each subject follows a linear regression model where some of the regression parameters are population specific(i.e. same for all subjects), whereas other parameters are subject-specific. From (3.4), the $\mathbf{b}_i$ are assumed to be random and are therefore often called random effects.
In general, a linear mixed effects model is any model which satisfies (Laird and Ware 1982)

$$\mathbf{Y}_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{3.6}$$

where $\mathbf{b}_i \sim N(0, D)$, $\epsilon_i \sim N(0, \sum_i), \mathbf{b}_1, \ldots, \mathbf{b}_N, \epsilon_1, \ldots, \epsilon_N$ independent,
where $\mathbf{Y}_i$ is the ni-dimensional response vector for subject $i, 1 \leq i \leq N$,N is the number of subjects, X_i and Z_i are $n_i\ x\ p$ and $n_i\ x\ q$ dimensional matrices of known covariates ,β is a p-dimensional vector containing the fixed effects,$\mathbf{b}_i$ is the q-dimensional vector containing the random effects,and ϵ_i is an ni-dimensional vector of residual components. Finally,D is a general$(q\ x\ q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$ and Σ_i is a $(n_i\ x\ n_i)$ covariance matrix which depends on i only through its dimension n_i,i.e. the set of unknown parameters in Σ_i will not depend upon i.
The importance of filtering out genes whose intensity is not well distinct from background level is supported by the following analytical arguments.
Let I_1 and I_2 be measured intensities for a given gene g in two different experimental conditions. Let β be the level of background intensity. Assume $I_1 \geq I_2 \geq 0$ and $0 \leq \beta \leq I_2$. Define the log-ratio of background subtracted intensities as :

$$log - ratio = log\left(\frac{I_1 - \beta}{I_2 - \beta}\right) \tag{3.7}$$

It will be convenient to think of the log-ratio as a function of β.
First of all, it is clear that as β approaches I_2 the value of log-ratio tends to infinity. More precisely, we can show that log - ratio is an increasing function of β. (and since $I_1 \geq I_2$,
Therefore, when $I_1 \geq I_2$.
In words, the derivative of log - ratio with respect to b is positive when $I_1 \geq I_2$. Therefore, log - ratio is an increasing function of β.
Now considering that log ....., we can conclude that the absolute value of the log - ratio is an increasing function of β, the background intensity level, irrespective of the relation between numerator and denominator.

Relationship between the log-ratio of background subtracted intensity and background intensity level. The x-axis is the value of the log-ratio of background corrected intensity. The measured intensity in the numerator is taken to be 600. The measured intensity in the denominator is taken to be 400.The x-axis is the value of background intensity.When the background intensity is very close to the gene intensity the log-ratio rapidly blows up to infinity. The x-axis is the value of β. The y- axis is the value of log- ratio . β is varied from 0 to I_2 .Finally, we choose arbitrarily $I_1 = 600$ and $I_2 = 400$, for purposes of illustration. We can see that when I_2 is near β, small changes in I_2 or in the background level get greatly exaggerated. This mathematical remark demonstrates that it is important to filter out genes whose intensity is not clearly above background level before performing background subtraction. Failing to do so can produce erroneous results in an analysis as simple as taking the log-ratio of two observations. We refer to the intensity that results from scanning the array and measuring intensity of florescence at each probe spot as "unprocessed intensity". Each unprocessed gene intensity observation is a measure of both the specific binding of a target mRNA on the probe and background labeling. An estimate of the local background intensity level is therefore subtracted from each unprocessed intensity to remove component that is not gene specific in the intensity observations

- Taking Logarithms :It will be important to transform DNA microarray data form the raw intensities into log intensities before proceeding with analysis mainly because

  a) There should be a reasonably even spread of features across the intensity range

  b) The variability should be constant at all intensity levels.

  c) The distribution of experimental errors should be approximately normal.

  d) The distribution of intensities should be approximately bell-shaped.

- Visualise the data with scatter plots and MA plots.

- Use within-array normalisation to remove effects of dye bias and spatial bias.

- Use between-array normalisation to enable comparison of multiple arrays.

Chapter 4

Results and discussion

4.1 Exploratory Data Anaysis

The primary goal of EDA is to maximize the our insight into the data set and into the underlying structure of the data set, while providing all of the specific items that we would want to extract from the data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings

boxplot
A boxplot displays the median, quartiles and extreme values of our gene expression data. So as to give a picture of the distribution of the data. From our data, it shows that Boran cattle had higher levels of gene expressions compared to Nda'ma. This is best explained by the fact

that since they are susceptible,there cells reacted more with the induced tryps. The results also shows us that Nda'ma cattle had more outliers as compared to Boran cattle. Boran cattle also had higher median as compared to Nda'ma.
Boxplot of the Boran and Ndama gene expressions

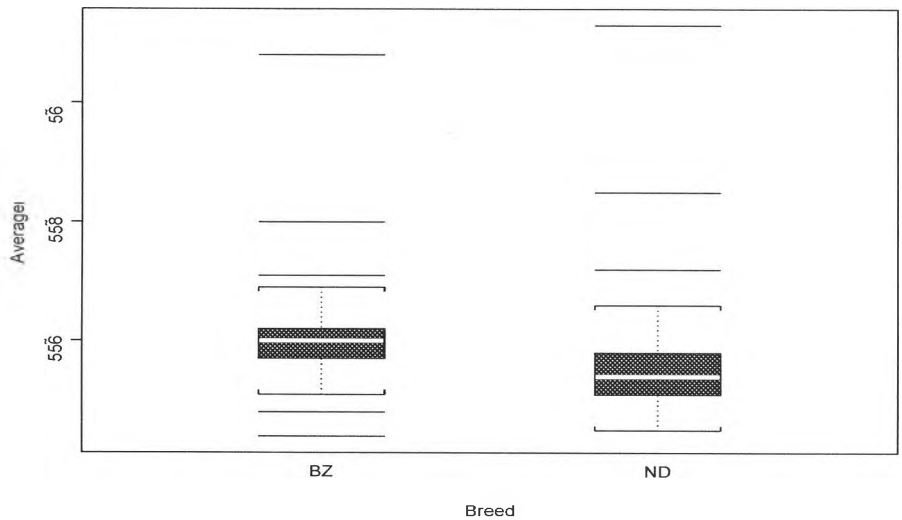

Figure 4.1: Liver box plot

4.2 Summary

Table 4.1: LIVER SUMMARY

Breed	Mean	Standard deviation	Median	N
Boran (BZ)	5.5607	0.00004854016	5.5600	58
Ndama (ND)	5.5559		5.5540	59

4.3 Analyses

4.3.1 Statistical model

Gene expressions were modeled via the hierachical linear model using S-Plus software. HLM is appropriate for data with a nested structure, sch as repeated-measures data in which several individual measurements(gene-expressions) are netsted or clustered within individuals (cattle). HLM separates within-cattle (Level 1) and between-cattle models (Level 2),estimating within-cattle and between-cattle variability:

$$Level - 1model, Y_{it} = \beta_{0i} + \beta_{1i}(T)_{it} + \beta_{xi}(X)_{it} + r_{it} \tag{4.1}$$

$$Level - 2model, \beta_{0i} = \gamma_{00} + \gamma_{01} Zi + U_{0i} \tag{4.2}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} Zi + U_{1i} \tag{4.3}$$

$$\beta_{xi} = \gamma_{x0} + \gamma_{11} Z_i \tag{4.4}$$

In the Level 1 model, Y_{it} is the gene expression for cattle i on a $particulardayt$ T is the group for a cattle i at a time t and X is a time-varying variable,i.e.,specific day. β_{0i} is the intercept indicating a cattle $i's$ fitted value of gene expression when both $TandX$ equal 0. $\beta_{1i}(T)_{it}$ is a slope, indicating cattle $i's$ gene expression per day. $\beta_{xi}(X)_{it}$ is a slope,indicating the association between time-varying variables and gene expressions. The random effect for the Level 1 model is given by r_{it}, and it is assumed to be normally distributed with mean 0 and variance $sigma^2$. The intercepts and slopes of the within-cattel model(Level 1) become the outcomes for the between cattle model (Level 2). The γ paramaters represent the mean level of the corresponding within-cattle parameters; γ_{00} corresponds to the mean initial status for gene expression; γ_{10} the average change per day; and γ_{x0} the average effect of time-varying variables on gene expressions. Any between-cattle variation in the regression coefficients is modelled in the Level 2 model as a function of a fixed factor Z_i and random effects U_{0i} and U_{1i}. These random effects are assumed to be normally distributed with means 0 and variances r_{00}^2 and r_{11}^2. Hierachical linear modelling allows us to

4.3.2 random intercept

From the table above, we conlude that the best model to adapt is the model with breed and day interaction since it has the lowest AIC (i.e 1378.341) value and the highest loglikehood value

Table 4.2: Model selection Deviance Table

Model Terms	Loglikelihood	Archaic Information Criterion	Bayesian Information Criterion
Breed	-687.9654	1383.931	1394.911
Day	-692.3233	1392.647	1403.626
Breed + Day	-684.8636	1379.727	1393.408
Breed * Day	-683.1707	1378.341	1394.706

(i.e -683.1707) which means that model with breed and day interaction best explain the gene expression results.

Table 4.3: Table of effects for breed with random Intercept

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Breed				
	Boran (BN)	Refference	——-	——-	——-
	Nd'ama(ND)	-24.2	8.44	-24.2 ± 16.54	0.008
Random part	Intercept				
	σ_b^2	38.07		$< .001, 10^8$	——
	σ_e^2	8569.21		78.20, 105.20	

From the above results,we conclude that the breed effect is very significant since the p-value 0.008 is far much less than 0.05 and therefore the different breeds express differently when induced with tryps over a period of time.

Intraclass correlation

The Intraclass Correlation is the proportion of variance that is accounted for by the group level. It is equal to the correlation between values of two randomly drawn micro-units in the same, randomly drawn, macro-unit. It assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. This will

help us to see how alike our geneexpression results are over time within the individual cattle and also within the groups of cattle
The theoretical formula for the ICC is:

$$\rho = \frac{\sigma^2 b}{\sigma^2 b + \sigma^2 e} \tag{4.5}$$

where $\sigma^2 e$ is the pooled variance within cattle, and $\sigma^2 b$ is the variance of the trait between cattle. It is easily shown that $\sigma^2 b + \sigma^2 e$ = the total variance of ratings–i.e., the variance for all ratings, regardless of whether they are for the same subject or not. Hence the interpretation of the ICC as the proportion of total variance accounted for by within-subject variation.
Therefore, from our results; our ICC for the model with breed alone with random intercept is;

$$\rho = \frac{6.17^2}{6.17^2 + 92.57^2} = 0.004 \tag{4.6}$$

we can therefore conclude that the between cattle variation is not significant since this value 0.004 explains only 0.4% of the between cattle variation.

Table 4.4: Table of effects for day with random intercept

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Day	1.52	0.65	1.52 ± 3.02	0.02
Random part	Intercept				
	σ_b^2	5.53^2			
	σ_e^2	90.57^2			

From the above results, we conclude that the day's effect is very significant since the p-value 0.02 less than 0.05 and therefore as the days increase or time changes the gene expression increases.
Therefore, from our results; our ICC for the model with breed alone with random intercept is;

$$\rho = \frac{5.53^2}{5.53^2 + 90.57^2} = 0.0037 \tag{4.7}$$

we can therefore conclude that between days variation is not significant since the ρ results is at 0.0037 which explains only 0.003 of the data variation.

Table 4.5: Table of effects for breed and day with random Intercept

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Breed				
	Boran (BN)	Refference	——-	——-	——
	Nd'ama(ND)	-24.2	8.44	-24.2 ± 16.54	0.007
Fixed part	Day	1.50	0.66	1.50 ± 1.29	0.02
Random part	Intercept				
	σ_b^2	36.60		$< .001, 10^8$	——
	σ_e^2	8226.49		78.20,105.20	

From the above results, it shows that as the days move or continue,theres a change in gene expression of one to two units.
Therefore, from our results; our ICC for the model with breed alone is;

$$\rho = \frac{6.05^2}{6.05^2 + 90.70^2} = 0.004 \tag{4.8}$$

we can therefore conclude that the between days variation is does not affect the changes in gene expression.
From the above results, it shows that the interaction of breed and day is not significant and therefore we conclude that the interaction of breed and day does not affect the change of gene expressions.

4.3.3 multilevel models i.e random slope

From the table above, we conlude that the best model to adapt to analyse the microarray data is the one with breed and day interaction since the AIC i.e(1376.672) is smaller compared to the one with random intercept and also the loglikelihood value i.e (-680.3362) is largest of all models.
From the above results, we conclude that the breed effect is very significant since the p-value 0.008 is far much less than 0.05 and therefore the different breeds express differently when induced with tryps over a period of time.

Table 4.6: Table of effects for the interaction breed and Day with random Intercept

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Breed				
	Boran (BN)	Refference	———	———	———
	Nd'ama(ND)	-13.84	13.35	-13.84 ± 26.20	0.3062
Fixed part	Day	2.16	0.92	2.16 ± 1.8	0.0218
Fixed part	Breed:day	-1.32	1.31	-1.32 ± 2.57	0.31
Random part	Intercept				
	σ_b^2	38.44			———
	σ_e^2	8259.17			

From the above results,we conclude that the day's effect is very significant since the p-value 0.02 less than 0.05 and therefore as the days increase or time changes the gene expression increases.
From the above results, it shows that as the days move or continue,theres a change in gene expression of one to two units.
From the above results, it shows that the ineraction of breed and day is not significant and therefore we conclude that the interaction of breed and day does not affect the change of gene expressions.

Table 4.7: Model selection Deviance Table

Model Terms	Loglikelihood	Archaic Information Criterion	Bayesian Information Criterion
Breed	-684.8879	1381.776	1398.245
Day	-688.6512	1389.302	1405.772
Breed + Day	-681.8786	1377.757	1396.911
Breed * Day	-680.3362	1376.672	1398.492

Table 4.8: Table of effects for breed with random slope

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Breed				
	Boran (BN)	Refference	——-	——-	——
	Nd'ama(ND)	-20.7	8.17	-20.7 ± 16.00	0.0152
Random part	Intercept				
	Day				
	σ_b^2	20.43			——
	σ_e^2	6446.48			

Table 4.9: Table of effects for Day with random slope

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Day	1.65	0.75	1.65 ± 1.47	0.03
Random part	Intercept				
	Day		2.77		
	σ_b^2	19.8			——
	σ_e^2	6494.75			

Table 4.10: Table of effects for breed and Day with random slope

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Breed				
	Boran (BN)	Refference	——-	——-	——
	Nd'ama(ND)	-20.8	8.17	-20.8 ± 16.00	0.0149
Fixed part	Day	1.64	0.74	1.64 ± 1.45	0.02
Random part	Intercept				
	Day		2.8		
	σ_b^2	20.07			——
	σ_e^2	6318.66			

Table 4.11: Table of effects for the interaction of breed and Day with random slope

Model part	Model terms	Parameter estimates	Standard error	95 % confidence interval	P-value P-value
Fixed part	Breed				
	Boran (BN)	Refference	——-	——-	——
	Nd'ama(ND)	-13.84	13.35	-13.84 ± 26.20	0.3062
Fixed part	Day	2.16	0.92	2.16 ± 1.80	0.0218
Fixed part	Breed:day	-1.32	1.31	-1.32 ± 2.57	0.31
Random part	Intercept				
	σ_b^2	38.44			——
	σ_e^2	8259.17			

Chapter 5

Conclusion

5.1 Summary and conclusion

From the above results, we can conlude that the best models to adapt to are the multilevel models (random slope) especially the model with breed and day interaction since it has the lowest AIC (i.e 1376.672) value and the largest loglikehood value (i.e -680.3362) which means that model with breed and day interaction best explain the gene expression results.
We can also conclude that the breed effect is very significant since the p-value 0.008 is far much less than 0.05 and therefore the different breeds express differently when induced with tryps over a period of time.
The day's effect is also very significant since the p-value 0.02 less than 0.05 and therefore as the days increase,move or time changes the gene expression increases.
The Intraclass correlation measures show us that between cattle variation is not significant since the ρ value 0.004 explains only 0.4% of the gene expressions.
The intraclass corellation between days variation is not significant since the ρ results is at 0.0037 which explains only 0.3% of the gene expressions, thus concluding that the between days variation does not affect the changes in gene expression.

5.2 references

1. Bates, D.M.(2007). Theory and Computational methods for mixed models. Department of statistics U. of Wisconsin-Madison

2. Bates, D.M.(2007). Linear mixed-effects models using S4 classes. U. of Wisconsin-Madison

3. Bates, D.M.(2007). Linear mixed model implementation in lme4. Department of statistics U. of Wisconsin-Madison

4. Bates, D.M.; R Development Core Team (2006). Statistical Inferences in Linear Mixed-Effects Models

5. Havatta, I. and Laine, M.M.(2005); DNA microarray and Data analysis.

6. Bates, D.M.(2004). Multilevel Models in R: Present and Future. U. of Wisconsin-Madison, U.S.A.

7. Bates, D.M. and DebRoy, S.(2004). Linear mixed models and penalized least squares. Journal of Multivariate Analysis.

8. Bates, D.M.; R Development Core Team (2004). Sparse Matrix Representation of Linear mixed models.

9. Alvarado, B.A. and Jairo, O.(2004). Growth trajectories are influenced by breast feeding and infant health in an afro-columbian community. Journal of Nutrition.

10. Friend, S.H. and Roland, B.S.(2002); The magic of microarrays.

11. Danh, V.N. and Arpat A.B.(2002); DNA microarray experiments: Biological technological aspects.

12. Snijders, A.B. and Bosker, R.J.(2002); An introduction to basic and advanced multilevel modelling.

13. Bates, D.M. and Pinheiro,J.C.(). Computational methods for Multilevel Modelling.

14. Laird, N.M. and Ware, J.H.(1982). Random-effects models for Longitudinal data. Biometrics 38, 963-974.

15. Lindstrom, M.J. and Bates, D.M.(1990). Nonlinear mixed effects models for repeated measures data. Biometrics 46, 673-687.

16. Lindstrom, M.J. and Bates, D.M.(1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects models for REpeated Measures Data. Journal of the American Statiatical Association 83, 1014-1022.

17. Pinheiro, J.C. and Bates, D.M.(2000). Mixed-Effects models in S and S-PLUS. Springer, 2000. ISBN 0-387-98957-9.

18. Pinheiro, J.C. and Bates, D.M.(1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-Effects model. Journal of Computational and Graphical Statistics 4, 12-35.

19. Pinheiro, J.C. and Bates, D.M.(1995). Mixed-Effects models methods and classes for S and Splus. U. of Wisconsin-Madison.

20. Pinheiro, J.C.(1994). "Topics in Mixed-Effects models." PhD thesis, U. of Wisconsin-Madison, Department of Statistics.

Appendix A

The Appendix

A.1 Summary statistics

```
liver1_importData("c:/cleaned_data/liver1.csv",type="ASCII")
temp_importData("c:/cleaned_data/liver1.xls",type="EXCEL")
temp1_t(temp)
#princomp(temp1)
attach(liver1)
#run sequential
plot (liver1,type="1")
#It appears that the data has fixed location and fixed scale
histogram(~average)
histogram(~Breed)
liver1_split(liver,Breed)
summary(liver1$BZ)
  AnimalNumber   Breed         Day       Tissue       Average
    Min.: 85.0   BZ:58      Min.: 0.00   Li:58       Min.:5.5440
 1st Qu.: 89.0   ND: 0   1st Qu.: 0.00             1st Qu.:5.5570
  Median:100.0            Median:18.00              Median:5.5600
    Mean:108.7              Mean:15.74                Mean:5.5607
  AnimalNumber   Breed         Day       Tissue       Average
 3rd Qu.:128.8           3rd Qu.:28.25             3rd Qu.:5.5620
    Max.:133.0              Max.:35.00                Max.:5.6080
summary(liver1$ND)
  AnimalNumber   Breed         Day       Tissue       Average
    Min.:171.0   BZ: 0      Min.: 0.00   Li:59       Min.:5.5450
 1st Qu.:175.5   ND:59   1st Qu.: 0.00             1st Qu.:5.5510
  Median:180.0            Median:18.00              Median:5.5540
    Mean:180.5              Mean:15.93                Mean:5.5559
  AnimalNumber   Breed         Day       Tissue       Average
 3rd Qu.:185.0           3rd Qu.:27.50             3rd Qu.:5.5580
    Max.:190.0              Max.:35.00                Max.:5.6130
```

A.2 Model Selection

A.2.1 Random intercept

```
> liver1 <- importData("c:/cleaned_data/liver1.csv", type = "ASCII")
> temp <- importData("c:/cleaned_data/liver1.xls", type = "EXCEL")
> temp1 <- t(temp)
> #princomp(temp1)
attach(liver1)
> #run sequential
#plot (liver1,type="1")
#It appears that the data has fixed location and fixed scale
#histogram(~average)
#histogram(~Breed)
liver1 <- split(liver, Breed)
> summary(liver1$BZ)
  AnimalNumber   Breed         Day       Tissue       Average
    Min.: 85.0   BZ:58      Min.: 0.00   Li:58       Min.:5.5440
 1st Qu.: 89.0   ND: 0   1st Qu.: 0.00            1st Qu.:5.5570
  Median:100.0            Median:18.00            Median:5.5600
    Mean:108.7              Mean:15.74              Mean:5.5607
 3rd Qu.:128.8           3rd Qu.:28.25           3rd Qu.:5.5620
    Max.:133.0              Max.:35.00              Max.:5.6080
> summary(liver1$ND)
  AnimalNumber   Breed         Day       Tissue       Average
    Min.:171.0   BZ: 0      Min.: 0.00   Li:59       Min.:5.5450
 1st Qu.:175.5   ND:59   1st Qu.: 0.00            1st Qu.:5.5510
  Median:180.0            Median:18.00            Median:5.5540
    Mean:180.5              Mean:15.93              Mean:5.5559
 3rd Qu.:185.0           3rd Qu.:27.50           3rd Qu.:5.5580
    Max.:190.0              Max.:35.00              Max.:5.6130
> xyplot(Average ~ Day | AnimalNumber, group = Breed, type = "b", panel = panel.superpose)
> xyplot(Average ~ Day | Breed, group = AnimalNumber, type = "b", panel = panel.superpose)
> options(contrasts = c("contr.treatment", "contr.poly"))
> fit1 <- lme(fixed = Average1 ~ Breed, random =  ~ 1 | AnimalNumber, liver)
> fit2 <- lme(fixed = Average1 ~ Day, random =  ~ 1 | AnimalNumber, liver)
> fit3 <- lme(fixed = Average1 ~ Breed + Day, random =  ~ 1 | AnimalNumber, liver)
> fit4 <- lme(fixed = Average1 ~ Breed * Day, random =  ~ 1 | AnimalNumber, liver)
> summary(fit1)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1383.931 1394.911 -687.9654

Random effects:
 Formula:  ~ 1 | AnimalNumber
        (Intercept) Residual
StdDev:   0.9186001 92.56763

Fixed effects: Average1 ~ Breed
```

```
               Value Std.Error DF   t-value p-value
(Intercept)  55607.24  12.15656 77  4574.257  <.0001
      Breed    -48.26  17.11900 38    -2.819  0.0076

Standardized Within-Group Residuals:
       Min         Q1        Med        Q3       Max
 -1.806556 -0.4214433 -0.205033 0.2270612 6.168089

Number of Observations: 117
Number of Groups: 40
> summary(fit2)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1392.647 1403.626 -692.3233

Random effects:
 Formula:  ~ 1 | AnimalNumber
        (Intercept) Residual
StdDev:    24.04969  90.5665

Fixed effects: Average1 ~ Day
               Value Std.Error DF  t-value p-value
(Intercept) 55559.15  13.84502 76 4012.933  <.0001
        Day     1.52   0.65450 76    2.316  0.0233

Standardized Within-Group Residuals:
       Min         Q1        Med        Q3       Max
 -1.373632 -0.5315192 -0.1556315 0.310579 5.531288

Number of Observations: 117
Number of Groups: 40
> summary(fit3)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1379.727 1393.408 -684.8636

Random effects:
 Formula:  ~ 1 | AnimalNumber
        (Intercept) Residual
StdDev:    5.996194  90.7003

Fixed effects: Average1 ~ Breed + Day
               Value Std.Error DF   t-value p-value
(Intercept)  55583.50  15.79365 76  3519.358  <.0001
      Breed    -48.46  16.88478 38    -2.870  0.0067
        Day      1.51   0.65333 76     2.306  0.0238

Standardized Within-Group Residuals:
       Min         Q1        Med        Q3       Max
```

```
 -1.576547 -0.4765251 -0.1652088 0.179849 6.053268

Number of Observations: 117
Number of Groups: 40
> summary(fit4)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1378.341 1394.706 -683.1707

Random effects:
 Formula:  ~ 1 | AnimalNumber
        (Intercept) Residual
StdDev:    2.367376 90.85779

Fixed effects: Average1 ~ Breed * Day
                Value Std.Error DF   t-value p-value
(Intercept)  55573.31  18.76765 75  2961.122  <.0001
      Breed    -27.68  26.69185 38    -1.037  0.3062
        Day      2.16   0.91976 75     2.344  0.0217
  Breed:Day     -1.32   1.30875 75    -1.006  0.3177

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3      Max
 -1.466329 -0.5004538 -0.1723792 0.2200243 6.160097

Number of Observations: 117
Number of Groups: 40
```

A.2.2 Random slope

```
> liver1 <- importData("c:/cleaned_data/liver1.csv", type = "ASCII")
> temp <- importData("c:/cleaned_data/liver1.xls", type = "EXCEL")
> temp1 <- t(temp)
> #princomp(temp1)
attach(liver1)
> #run sequential
#plot (liver1,type="1")
#It appears that the data has fixed location and fixed scale
#histogram(~average)
#histogram(~Breed)
liver1 <- split(liver, Breed)
> summary(liver1$BZ)
  AnimalNumber   Breed         Day       Tissue       Average
    Min.: 85.0   BZ:58      Min.: 0.00   Li:58       Min.:5.5440
 1st Qu.: 89.0   ND: 0   1st Qu.: 0.00            1st Qu.:5.5570
  Median:100.0            Median:18.00             Median:5.5600
    Mean:108.7              Mean:15.74               Mean:5.5607
 3rd Qu.:128.8           3rd Qu.:28.25            3rd Qu.:5.5620
    Max.:133.0              Max.:35.00               Max.:5.6080
```

```
> summary(liver1$ND)
  AnimalNumber   Breed         Day        Tissue       Average
    Min.:171.0   BZ: 0      Min.: 0.00   Li:59       Min.:5.5450
 1st Qu.:175.5   ND:59   1st Qu.: 0.00            1st Qu.:5.5510
  Median:180.0            Median:18.00            Median:5.5540
    Mean:180.5              Mean:15.93              Mean:5.5559
 3rd Qu.:185.0           3rd Qu.:27.50           3rd Qu.:5.5580
    Max.:190.0              Max.:35.00              Max.:5.6130
> xyplot(Average ~ Day | AnimalNumber, group = Breed, type = "b", panel = panel.superpose)
> xyplot(Average ~ Day | Breed, group = AnimalNumber, type = "b", panel = panel.superpose)
> options(contrasts = c("contr.treatment", "contr.poly"))
> fit1b <- lme(fixed = Average1 ~ Breed, random =  ~ Day | AnimalNumber, liver)
> fit2b <- lme(fixed = Average1 ~ Day, random =  ~ Day | AnimalNumber, liver)
> fit3b <- lme(fixed = Average1 ~ Breed + Day, random =  ~ Day | AnimalNumber, liver)
> fit4b <- lme(fixed = Average1 ~ Breed * Day, random =  ~ Day | AnimalNumber, liver)
> summary(fit1b)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1381.776 1398.245 -684.8879

Random effects:
 Formula:  ~ Day | AnimalNumber
 Structure: General positive-definite
               StdDev   Corr
(Intercept) 21.580217 (Inter
        Day  3.050772 -1
   Residual 80.292461

Fixed effects: Average1 ~ Breed
                Value Std.Error DF   t-value p-value
(Intercept)  55599.00  11.56279 77  4808.441  <.0001
      Breed    -41.56  16.33935 38    -2.544  0.0152

Standardized Within-Group Residuals:
       Min         Q1        Med        Q3      Max
 -1.830298 -0.4249589 -0.1064738 0.2464359 4.519972

Number of Observations: 117
Number of Groups: 40
> summary(fit2b)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1389.302 1405.772 -688.6512

Random effects:
 Formula:  ~ Day | AnimalNumber
 Structure: General positive-definite
               StdDev   Corr
(Intercept) 11.244653 (Inter
```

```
        Day  2.772276 -1
   Residual 80.593329

Fixed effects: Average1 ~ Day
               Value Std.Error DF  t-value p-value
(Intercept) 55558.42  12.02108 76 4621.749  <.0001
        Day     1.65   0.74625 76    2.210  0.0301

Standardized Within-Group Residuals:
       Min         Q1         Med        Q3      Max
 -1.503712 -0.4869644 -0.09053924 0.3110786 4.445019

Number of Observations: 117
Number of Groups: 40
> summary(fit3b)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1377.757 1396.911 -681.8786

Random effects:
 Formula:  ~ Day | AnimalNumber
 Structure: General positive-definite
               StdDev   Corr
(Intercept) 16.982229 (Inter
        Day  2.800248 -1
   Residual 79.492628

Fixed effects: Average1 ~ Breed + Day
                Value Std.Error DF   t-value p-value
(Intercept)  55579.30  14.52951 76  3825.270  <.0001
      Breed    -41.67  16.34429 38    -2.550  0.0149
        Day      1.64   0.74166 76     2.218  0.0296

Standardized Within-Group Residuals:
       Min        Q1        Med       Q3     Max
 -1.684671 -0.452028 -0.1364447 0.231894 4.48228

Number of Observations: 117
Number of Groups: 40
> summary(fit4b)
Linear mixed-effects model fit by REML
 Data: liver
       AIC      BIC    logLik
  1376.672 1398.492 -680.3362

Random effects:
 Formula:  ~ Day | AnimalNumber
 Structure: General positive-definite
               StdDev   Corr
(Intercept) 17.748052 (Inter
```

```
       Day  2.826307 -1
  Residual 79.701451

Fixed effects: Average1 ~ Breed * Day
               Value Std.Error DF   t-value p-value
(Intercept)  55573.45  17.00898 75  3267.300  <.0001
      Breed    -29.75  24.18237 38    -1.230  0.2261
        Day      2.14   1.04777 75     2.042  0.0447
  Breed:Day     -1.00   1.49089 75    -0.672  0.5035

Standardized Within-Group Residuals:
       Min         Q1        Med        Q3      Max
-1.615519 -0.4685422 -0.1296131 0.2763263 4.450552

Number of Observations: 117
Number of Groups: 40
```