Microarray Experiments to estimate Heterosis:

Design, Transformations, Models

DISSERTATION

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund
vorgelegt von

Barbara Sarholz

Hohenheim 2007

Prüfungskommission:	Prof. Dr. Götz Trenkler (Vorsitzender)
	Prof. Dr. Hans-Peter Piepho (Gutachter)
	Prof. Dr. Jörg Rahnenführer (Gutachter)
	Prof. Dr. Wolfgang Urfer (Gutachter)
	Dr. Holger Schwender (Beisitzer)
Tag der mündlichen Prüfung:	28. September 2007

Danksagung

Während meiner Arbeit an der Universität Hohenheim innerhalb des von der Deutschen Forschungsgemeinschaft (DFG) finanzierten Schwerpunktprogramms ″Heterosis in Pflanzen″ konnte ich nun die vorliegende Dissertation abschließen. Durch diese für mich wichtige Zeit haben mich einige Menschen begleitet, denen ich für ihre Unterstützung danken möchte.

An erster Stelle danke ich Herrn Piepho für die erstklassige Betreuung. Er ließ mir einerseits großen Freiraum, andererseits nahm er sich viel Zeit, wenn Unterstützung gefragt war. Seine konstruktive Kritik half mir, meine Arbeit in vielen Punkten gezielt zu verbessern.

Herr Rahnenführer und Herr Urfer erklärten sich sofort bereit, die Arbeit zu begutachten und haben durch ihren kritischen Blick zum Gelingen der Dissertation beigetragen. Auch ihnen sei an dieser Stelle herzlich gedankt.

Des weiteren danke ich den Partnern des DFG-Projektes ″Heterosis in Pflanzen″, insbesondere Stephanie Meyer und Stefan Scholten von der Universität Hamburg sowie Frank Hochholdinger und Nadine Höcker von der Universität Tübingen. In gemeinschaftlichen Diskussionen tauchten Fragestellungen auf, aus denen schließlich die vorliegende Dissertation entstanden ist. Sie stellten mir ihre Datensätze zur Verfügung,

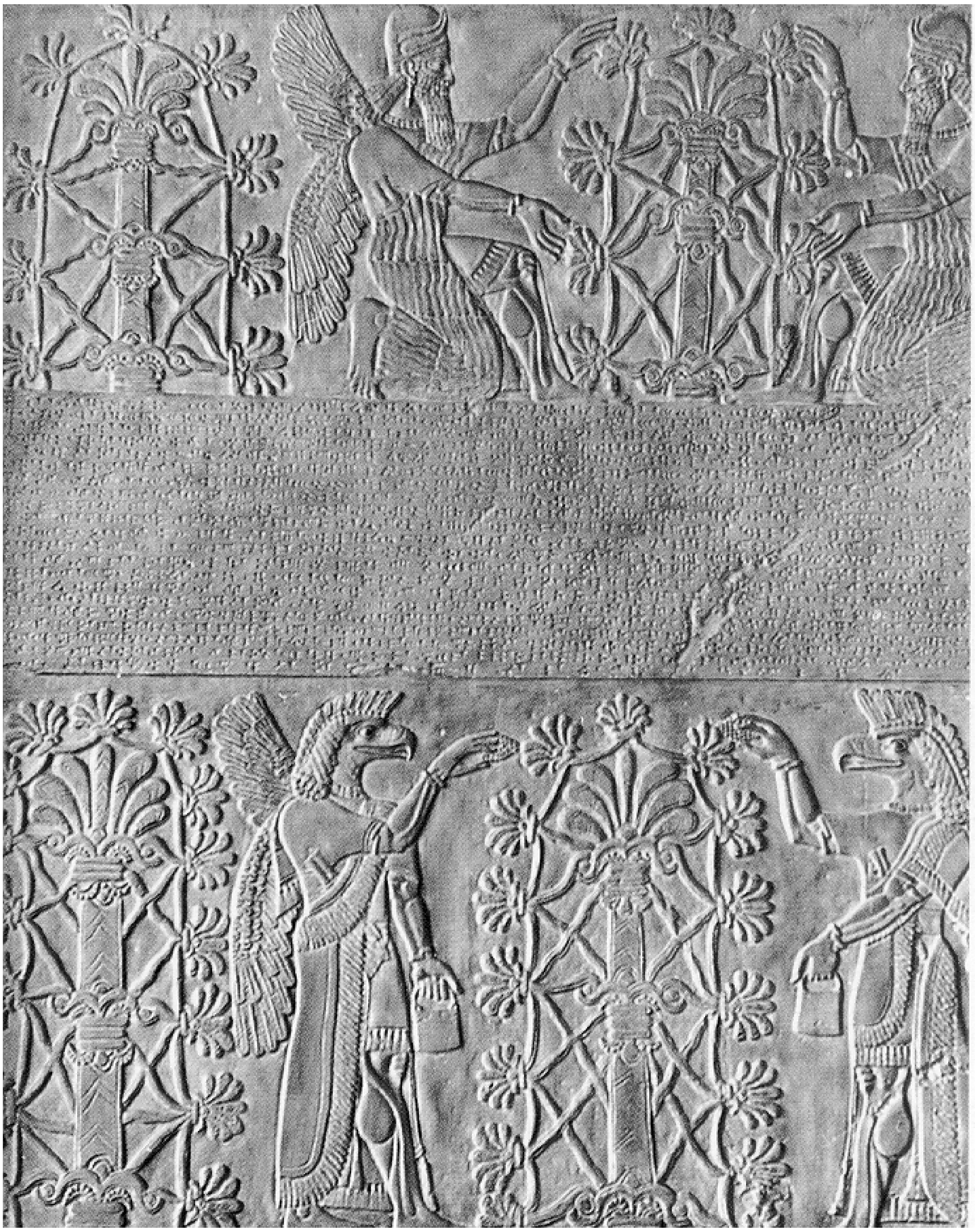
wodurch die Arbeit sehr praxisnah gehalten werden konnte.

Neben den externen Projektpartnern möchte ich meine Kollegen am Institut für Pflanzenbau und Grünland nicht unerwähnt lassen. Andreas Büchse, Katharina Emrich, Karin Hartung, Anita Kämpf, Irina Kuzyakova, Martina Mayus, Jens Möhring, Bettina Müller, Joseph Ogutu, Andrea Richter, André Schützenmeister, und Albrecht Weber trugen zur positiven Atmosphäre am Institut bei und sorgten dafür, dass es neben der Arbeit auch viel zu lachen gab.

Vielen Dank auch meinen Dortmunder Statistik-Freunden. Bei Fragen organisatorischer Natur fand sich stets ein kompetenter Ansprechpartner, und der Austausch von Fachartikeln machte die ein oder andere Fernleihbestellung überflüssig.

Zu guter Letzt gilt mein besonderer Dank Gisela Trabert und Tobias Sarholz, die alle Höhen und Tiefen der letzten Jahre miterlebt haben, die für Ablenkung sorgten, wenn die Zeit reif dafür war, und die mich im entscheidenden Moment wieder motivierten.

Barbara Keller (jetzt Sarholz) Hohenheim, im Juli 2007



Assyrians artificially pollinating date palms. The relief dates from between 883 and 859 B.C. (Mold at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany. Photograph: IPK Archive, Gatersleben)

Contents

Abstract

The genetic causes for heterosis, i.e., the increased performance of a hybrid plant compared to the parental mean, may be assessed via microarrays. This thesis addresses design and analysis issues of cDNA-microarray experiments with regard to the estimation of heterosis. Standard microarray designs like the loop design or common reference design are not optimal when estimating heterosis. An optimality criterion is devised and two approaches to obtain a suitable design are shown: a rather intuitive one and an approach using simulated annealing. Data transformations are crucial before analysing microarray data. However, transformations may conceal interesting expression patterns. It is shown using a Box-Cox transformation that significance of a heterotic effect is largely influenced by the transformation parameter. Transformation of the linear predictor in a generalized linear model has a similar effect and heterotic effects may—at least partially—be removed by the transformation. For the estimation of linear contrasts between genotypes, a linear mixed model for each gene is fitted to the expression values. To improve variance estimates one may benefit from other genes' information. Therefore, an empirical Bayes approach is developed that is capable of including more than one variance component in the model.

Zusammenfassung

Die genetischen Gründe für Heterosis, d.h. die erhöhte Leistung von Hybridpflanzen gegenüber dem Elternmittel, können mit Hilfe von Microarrays untersucht werden. Diese Doktorarbeit befasst sich mit Aspekten des Designs und der Analyse von cDNA-Microarrays im Hinblick auf die Schätzung von Heterosis. Standard-Microarraydesigns wie Loop- und Common-Reference-Design sind für die Schätzung von Heterosis nicht optimal. Um für die Heterosisschätzung geeignete Designs zu finden, wird ein Optimalitätskriterium entwickelt und zwei Ansätze zur Designsuche werden gezeigt: ein eher intuitiver Ansatz und einer, der die Methode des Simulated Annealing nutzt. Vor der Analyse von Microarray-Daten ist es meist notwendig, die Daten zu transformieren. Allerdings können Transformationen interessante Expressionsmuster verschleiern. Anhand der Box-Cox-Transformation wird gezeigt, dass die Signifikanz heterotischer Effekte stark vom Transformationsparameter abhängt. Die Transformation des linearen Prädiktors im Rahmen eines generalisierten linearen Modells hat einen ähnlichen Effekt und heterotische Effekte können, zumindest teilweise, durch die Transformation beseitigt werden. Dies sollte bei der Analyse berücksichtigt werden. Für die Schätzung linearer Kontraste von Genotypen wird ein gemischtes lineares Modell an die Genexpressionswerte angepasst. Um Varianzschätzungen zu verbessern,

kann die Information anderer Gene genutzt werden. Dazu wurde eine Empirical Bayes Methode entwickelt, bei der Informationen aller Gene genutzt werden und gleichzeitig die Einbeziehung mehrerer zufälliger Effekte ins Modell möglich ist.

Chapter 1

Introduction

For a long time people have known about heredity and that certain traits pass from one generation to the next. Evidence may be found in an Assyrian relief showing gardeners artificially pollinating date palms (see picture at the beginning of the thesis), which dates from between 883 and 859 B.C. Most of the crops we use today evolved after centuries of plant breeding. In the course of time more and more mechanisms of heredity were discovered, e.g., Mendel's laws or the double helix. However, there are still many mysteries left to solve. One of these is heterosis, the phenomenon that the crosses of two genetically distinct inbred lines, so-called hybrids, show better performance in many agronomic traits than their parents. Many of the various facets of heterosis are highly appreciated in plant breeding. Among the most important characteristics of hybrids are the increased yield and higher resistance against drought or pathogens compared to their parents. Furthermore, the performance of hybrids is more stable. Heterozygous plants are less subject to genotype-environment interactions and thus show improved reliability of yield (Léon, 1994; Becker, 1993). These advantages led to an increase in hybrid cultivation despite the higher effort in breeding and seed production. Today, many species such as maize, sugar-beet, and rye are cultivated wholly or predominantly with hybrids.

Although the benefits of heterosis have been applied for quite a while, the genetic and molecular causes are so far not fully understood. It was not until the middle of the last century that Avery found the DNA to be the bearer of genetic information. Since then, genetics evolved rapidly and nowadays new technologies allow us further insights into the genome and its functionality.

With the aid of microarrays, developed in the early 1990s, the expression of thousands of genes may be determined by a single experiment (Schena, 2003). By applying DNA of hybrids and parents, the difference in expression between the genotypes may be measured for each gene. Thus, microarrays seem to be a valuable tool in the exploration of heterosis. Research in this field seems to be worthwhile as these insights could be used to develop new strategies for plant breeding. In 2003 a Priority Program 'Heterosis in plants' (SPP 1149) was established, which is funded by the DFG (Deutsche Forschungsgemeinschaft). The objective of the program is the search for the molecular and genetic reasons of heterosis. For the experiments described in this study maize is used, as it shows very intense heterotic effects, e.g., for grain yield or plant height (Becker, 1993). Maize plants are supposed to have model character, i.e., the results may be carried forward to other species. This thesis emerged during my work in the project group 'Bioinformatic Tools for Microarrays' within the DFG Priority Program.

The following sections of this introduction will explain the fundamentals of both heterosis and microarray technology in greater detail.

1.1 Heterosis and dominance

When crossed, two genetically different inbred lines result in a heterozygous offspring, which is called hybrid or F_1 hybrid. The F_1 stands for 'first filial generation'. The increase in performance of the hybrid over their parental lines is called hybrid vigour or heterosis. It has been utilized in plant breeding since the middle of the 18th century, but a theory has never been formulated until the work of Shull (1908). The term heterosis was first used by Shull in 1917 during a lecture he gave in Goettingen. Up to now, underlying mechanisms of heterosis are not yet fully understood on the genetic and molecular level. Several quantitative genetic explanations that make the combination of a considerable number of genes responsible for heterosis have been discussed (for review see: Lamkey and Edwards (1998); Stuber (1999)) but little consensus has emerged. Most hypotheses were formulated before the molecular concepts of genetics were discovered and are not related to molecular principles (Birchler, Auger, & Riddle, 2003).

Heterosis may already be observed in early developmental stages as Höcker, Keller, Piepho, and Hochholdinger (2006) showed with early maize roots. However, the highest degree of heterosis is observed in agronomic traits of fully grown plants. In some types of cereal like maize and rye the heterosis-effect may double the yield of the hybrid compared to the parental inbred lines. Accordingly, the use of hybrids in crop production increased immensely during the last decades. Random mating of the F_1 in subsequent generations, i.e., crossing two hybrids from the same filial generation, usually leads to a reduced mean performance (Figure 1.1, found in Graw (2006)). This so-called inbreeding depression was already found by Darwin (1876). Since then, the genetic basis of heterosis has been

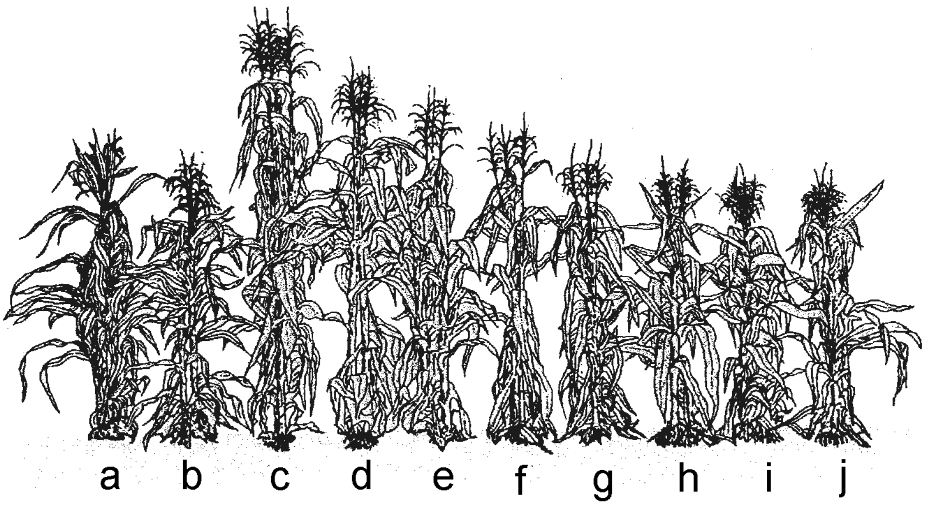


Figure 1.1: *Heterosis in maize. The hybrid (c) shows stronger growth compared to the parents (a and b). Later inbreeding generations reveal clearly reduced yield (d - j).*

discussed (Shull, 1908; East, 1908). Which genes exactly are responsible for heterotic effects and how they work together is yet unknown. With the rise in molecular biology during the last decades, new opportunities for the exploration of heterosis arise.

Let us turn towards the mathematical definition of heterosis. Let κ_A denote the expected value of a characteristic of line A, such as height or vigour; κ_B and κ_{AB} denote the same expectations for line B and hybrid AB, respectively. Heterosis is defined as the difference in performance of the hybrid compared to the mid-parent value, or, in mathematical terms

$$\delta(AB) = \kappa_{AB} - \frac{\kappa_A + \kappa_B}{2}. \tag{1.1}$$

$\delta(\cdot)$ is also denoted 'mid-parent heterosis' (MPH), contrary to the better-parent heterosis (BPH), which is defined as

$$\delta^*(AB) = \kappa_{AB} - \max(\kappa_A, \kappa_B). \tag{1.2}$$

Of course, not all hybrids show an increase in performance for all phenotypic characteristics; some hybrids may show equal or even inferior performance compared to the parents. In this study, however, a negative difference between the hybrid and the parental mean will also be denoted by heterosis.

If we have a closer look at a plant, its genome consists of thousands of genes, which are all composed of four nucleic acids. One may measure the expression level of a certain gene, i.e., the amount of mRNA. If we carry the definition of heterosis to the molecular level, 'heterosis' occurs when the expression level of a gene in a hybrid differs from the mean expression level of the parents. This phenomenon we denote dominance. In (1.1) κ then is the expression level of a defined gene. We use the term dominance in place of heterosis because dominance commonly refers to gene effects, while heterosis is usually defined in terms of phenotypic means for polygenic traits (Falconer & Mackay, 1996). Dominance may occur at various intensities: the expression level of the hybrid may lie between the expression levels of the parents (partial dominance), or the expression level of the hybrid may exceed that of both parents (overdominance). If the expression level of the hybrid is lower than the mid-parent level, we denote this as negative dominance. It is supposed that analysis of genes showing dominance in certain patterns will give a clue about how the phenomenon heterosis works.

1.2 Microarray technology

In the 1990s a technique was established that allows the simultaneous transcriptome-wide expression profiling of thousands of different genes in a single experiment. The so-called microarrays may be classified into oligonucleotide-arrays and two-channel cDNA arrays. Oligonucleotide-arrays were developed by the Affymetrix company. A gene is represented by 20 to $\sim$ 80 oligonucleotides (oligos). They are designed in such a way as to hybridise to different regions of RNA corresponding to an expressed gene.

In this thesis we will consider only cDNA-arrays. These contain a collection of cDNA spots, so called 'targets', that are attached to a small glass slide. To make sure that the spots adhere to the array, an electrically charged substrate is applied on the glass slide before spotting (Figure 1.2, found in Schena (2003)). The two test samples containing DNA of two different tissue types are called 'probes' and are marked with fluorescent

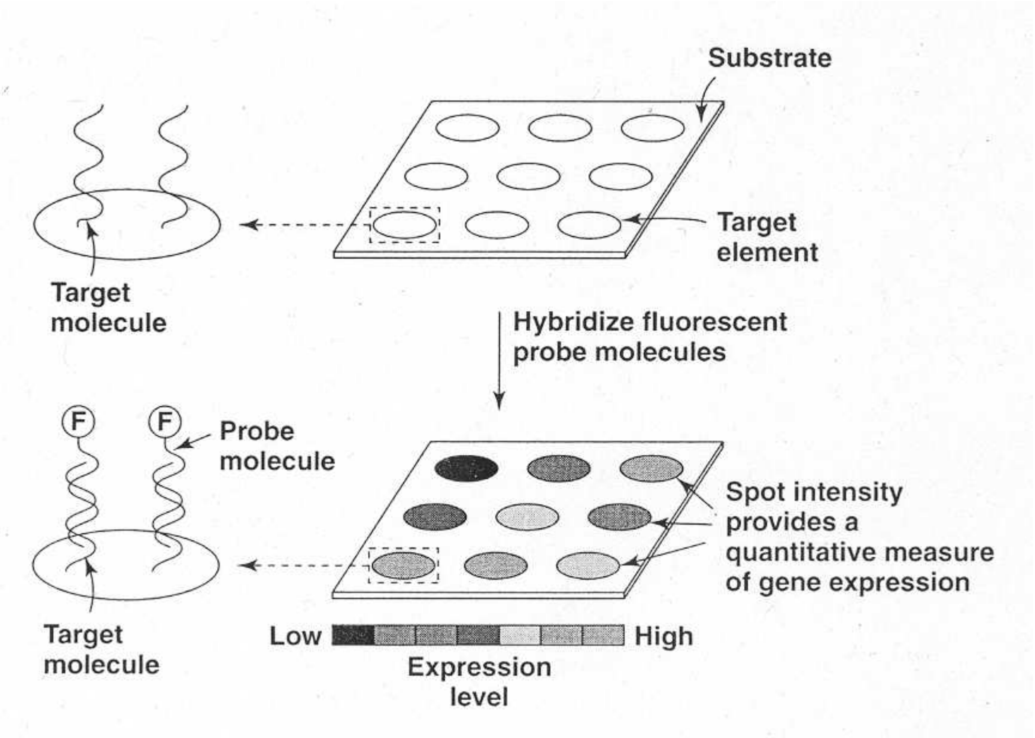

Substrate
Target element
Target molecule
Hybridize fluorescent probe molecules
F
F
Probe molecule
Spot intensity provides a quantitative measure of gene expression
Target molecule
Low
High
Expression level

Figure 1.2: *Microarray hybridisation*

dyes. Usually for this purpose Cy3 and Cy5 are used, two dyes that are excited by a green and red laser, respectively. In a hybridisation reaction the DNA of the test samples will bind to the cDNA on the array. Depending on the amount of corresponding DNA that is in a test sample, the spot on the array will appear more or less bright. The signal intensity therefore provides a quantitative measure of gene expression. The position and intensity of the spots are then detected by a laser scanner at two wavelengths, for the red and the green colours. Figure 1.3 shows the overlay of the two scans. Spots which are red or green correspond to genes which are mainly expressed in one of the two test samples. If a gene is expressed in none of them, it appears as a dark spot on the array, whereas a gene that is expressed in both samples appears yellow.

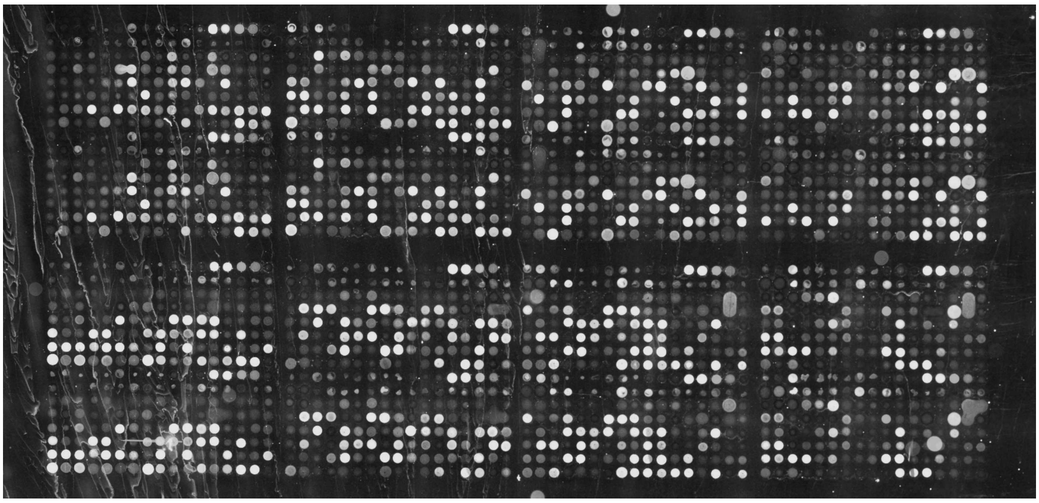


Figure 1.3: *Extract of microarray with maize genotypes UH005×UH301 and UH301*. S. Scholten, University of Hamburg.

In a step called image analysis the image produced by the scanner is converted into numerical information. There is a variety of different computer algorithms for this purpose implemented in software packages. For further information see Stekel (2003).

With the help of microarrays, global patterns of gene expression can be analysed at a defined developmental stage between different genotypes. A number of studies aiming to locate differentially expressed genes between inbred lines and reciprocal hybrid have been published (Ni, Sun, Liu, Wu, & Wang, 2000; Kollipara, Saab, Wych, Lauer, & Singletary, 2002; Guo, Rupe, Danilevskaya, Yang, & Hu, 2003) and the phenomenon of heterosis in maize is discussed in Auger et al. (2005), yet without the aid of microarrays.

Microarray data are highly noisy. This is partly due to the small size of the microarray, which is often no bigger than 2×4 cm^2, and the technically based inaccuracies resulting thereof. Often there are spatial effects on the array, as it is impossible to apply the probe mixture in a totally even manner. The targets are spotted by a robot onto the slides, and there may be irregularities of the printtips, possibly affecting the uniformity of the spots. The fluorescent dyes in general do not bind equally well to the probes, therefore one probe results in higher signal values, which also has to be accounted for. Due to the many sources of variation it is inevitable to perform replicated experiments. Signal values from different arrays may show considerable differences in both location and scale. All these sources of error are accounted for in the normalization step (Schena, 2003; Y. H. Yang, Dudoit, Luu, & Speed, 2001) described in Section 2.3.

The main purpose of most microarray experiments is the detection of differentially expressed genes. In a simple case when only two tissue types are investigated, the difference in expression of the tissues may be analysed with a t-test. With more sophisticated objectives and complex designs involving multiple sources of error, it is advisable to apply a linear model, e.g., when estimating heterosis contrasts with gene expression data from three genotypes. Other sources of variation than those resulting of the tissue types can be taken into account. Kerr and Churchill (2001) have noticed four main sources of variation: the tissue types, the fluorescent dyes used to label tissues, the genes (according to the spots on the array) and the different arrays used in the experiment. After performing the tests, the resulting p-values are usually adjusted for the multiplicity problem. As the arrays contain several thousand spots that are analysed separately, there are presumably many false positives. For the adjustment different methods are proposed, for example the control of the familywise error rate, the false positive rate (Hsueh, Chen, & Kodell, 2003), or the false discovery rate (Benjamini & Hochberg, 1995).

1.3 Outline

The thesis emerged out of practical problems, either when planning or analysing experiments aiming at the exploration of heterosis. Therefore, the following chapter is dedicated to the description of these experiments and data sets.

Chapter 3 is concerned with the first out of three statistical fields of interest related to heterosis covered by the thesis, namely the design of a microarray experiment. When investigating the developments for optimal microarray designs, the impression emerges that it was sometimes tried to reinvent the wheel. Microarray designs, however, may be considered as special cases of usual designs. A 'traditional' procedure of finding optimal designs when no analytical results are available is the numerical search. This should not be ignored here, especially as the finding of heterosis-relevant genes is an objective that is, in terms of statistics, different from the simple comparison of two genotypes. Therefore it cannot be handled in an optimal way with the classical microarray designs. We present two approaches that are distinguished from other microarray designs in the following points: The optimality criterion is tailored precisely to the objective of the study. It may account for an arbitrary number of effects, among those nuisance effects caused by technical reasons. Finally it is shown that information of earlier experiments may be reasonably utilised.

Chapter 4 throws a light on data transformations used to assess heterosis or dominance. Phenotypic data as well as gene expression data for heterosis estimation often lack validity of assumptions such as normality or homogeneity of variance, as do gene expression data derived from microarrays. If the data is transformed, or if a transformation is performed via a generalized linear model, these transformations have an impact on the heterosis or dominance estimate, which is easily ignored.

Chapter 5 deals with the estimation of variance components in microarray analysis. Due to the high costs of microarray experiments the number of replicates is usually quite low. However, as the analysis is done per gene, information on variance components may be shared across genes by specifying their distribution across genes. It is shown how pooled variance estimates may be determined using an empirical Bayes approach. The advantage of the approach is that the analysis may be performed with a model including more than one variance component, which is especially worthy when investigating heterosis. As the distribution is fitted to the sum of squares instead of the actual variance components, the problem of fitting a distribution to zeros is avoided, as sum of squares will be positive with probability one, provided the data have a multivariate normal distribution. Application to real microarray data shows that the approach supplies good results and is computationably feasible.

Finally, the last chapter contains an overview and a general discussion of the findings.

Chapter 2

Data, preprocessing and software

Trying to unravel the genetic causes for heterosis, the statistician is confronted with problems that were often neglected in microarray literature. Due to extensive cooperations within the DFG project ′Heterosis in plants′ with the Universities of Hamburg, Munich and Tübingen we were confronted with various real life tasks and had several data sets at our disposal. These served both as inspiration for research and as test data sets to validate assumptions and ensure that methods work in real-world problems. The experiments are all related to maize and are either phenotypic or microarray experiments. We now present the data sets that are referred to in the following chapters. To get a quick overview the data sets and corresponding chapters are indicated in Table 2.1.

Sections about data pre-processing and the applied software follow the description of the data sets.

Table 2.1: *Overview of data sets and the chapter where they are analysed*

data set		chapter
2.1.1	Heterosis microarray experiment	3
2.1.2	Microarray experiment wild type vs. mutant	3
2.1.3	Primary root length	3
2.1.4	Lateral root length	4
2.2	Hamburg data	5

2.1 The data from Tübingen

The work group General Genetics of the Center for Plant Molecular Biology, University of Tübingen (led by Frank Hochholdinger), is engaged in the exploration of early maize root development. The roots of the plant play an important role in water and nutrient supply. Figure 2.1 shows different root types of a maize plant in the seedling stage.

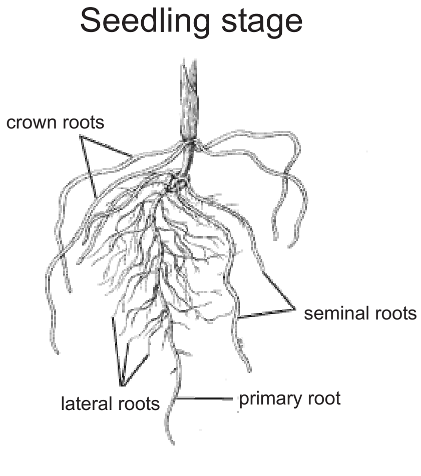


Figure 2.1: *Root types of maize, drawings by: Miwa Kojima, Schnable laboratory, Iowa State University*

2.1.1 Heterosis microarray experiment

This is the sole experiment where our focus lies on the design of experiment (although the data that arose has in the meantime been analysed by us), while for the other experiments we focus on data analysis. The planned experiment was performed in order to identify genes for which the expression level of the hybrid significantly exceeds the mean expression level of the parents. These genes will then be subjected to a subsequent detailed analysis. The experiment comprises altogether 16 inbred lines and F_1 hybrids. The chosen parental inbred lines are denoted with A (UH002), B (UH005), C (UH250) and D (UH301). All in all there are 12 hybrids, including reciprocal hybrids. The reciprocal of a hybrid is defined as a cross of the same parents, where the male and female parents are exchanged. The resulting hybrids are denoted as AB, AC, AD, BC, BD, CD and their reciprocals as BA, CA, DA, CB, DB, DC. For these hybrids and reciprocal hybrids a design is searched for the estimation of dominance, having a total of 72 arrays on-hand.

The objective of this microarray experiment was to find genes that show dominance effects (see Section 1.1). Therefore, the lines UH002, UH005, UH250, and UH301 as well as the hybrids and reciprocal hybrids were chosen. Our task was the development of an experimental design to determine dominance effects with high precision, using a total of 72 arrays.

2.1.2 Microarray experiment wild type vs. mutant

Usually, all maize seedlings develop a root system like that in Figure 2.1. However, there exists a mutant which does not form any crown- and lateral roots. The mutant was originally found in line DK105 but was since then crossed several times with B73. In order to find differences in gene expression between the mutant and the wild type B73, a microarray experiment was performed. On each of four replicates wild type and mutant were hybridised and a dye-swap was included.

2.1.3 Primary root length

This experiment was conducted to see if maize plants grown in different experimental units show phenotypic differences in their early root development. Two maize seeds were cultivated on filter papers. 16 filter papers were put together in one beaker. The experiment was conducted on two days, each day 7 beakers with maize plants were cultivated. Four days after germination the primary root length was determined. To keep genotype-environment interactions low, a hybrid (UH005 × UH301) was chosen instead of an inbred line.

2.1.4 Lateral root length

To investigate if different genotypes show differing lengths of lateral roots, maize seeds of the inbred lines UH005, UH250 and UH301 and the six resulting hybrids were cultivated under laboratory conditions. Ten days after germination a root zone of length 2-3 cm was cut approximately 20 cm distant from the root tip and the length of all lateral roots was measured. Of each genotype, between 8 and 21 primary roots were available. The number of lateral roots per primary root differed between 2 and 41. As we have only phenotypic data, effects cannot be ascribed to individual genes.

2.2 The data from Hamburg

A work group of the Department of Developmental Biology at the Biocenter Klein Flottbek, University of Hamburg (led by Stefan Scholten), studies the difference in gene expression between inbred lines and hybrids in early developmental stages of maize. About six days after pollination mRNA is extracted from embryo and endosperm, which are analysed in separate microarray experiments. Experiments were conducted with different genotypes. For this study, the genotypes UH250, UH301, UH250xUH301, and UH301xUH250 were chosen. Arrays were hybridised with all hybrid-parent combinations in three replicates, thus resulting in 12 arrays for embryo and 12 arrays for endosperm.

2.3 Data pre-processing

To eliminate sources of variation due to technical reasons it is necessary to perform a data normalization (Schena, 2003; Y. H. Yang et al., 2001). This does not lie within the main focus of the thesis but is described here shortly for completeness. The normalization procedure was performed with the data of all microarray experiments mentioned in the previous sections. The raw data comprises foreground and background expression values for each channel (Cy3, Cy5) and each spot. The foreground value is a measure of signal intensity of the actual spot, while the background value gives the intensity of the spot's surrounding. To account for unspecific background noise, the background value is subtracted from the signal value. A loess-normalizaton is then performed to $\log_2$ transformed data of each array. The loess-normalization accounts for intensity-based dye effects. As values from different arrays may differ considerably, median absolute deviations of each array's channel are adjusted. Thereafter, a linear mixed model is fitted to normalized data which will be described in the corresponding chapters (i.e., Chapter 3.2 and Chapter 5.3).

2.4 Software

The data analysis of this dissertation was performed using the packages STAT, GRAPH and IML of SAS® system software. For Chapter 3, Version 8 was applied, while the other chapters are based on Version 9 for Windows (SAS Institute Inc., 1999/ 2002-2003).

Chapter 3

Microarray design for the estimation of dominance effects

Many principles of experimental design were developed in the 1920s and 1930s by R. A. Fisher and F. Yates. The main applications were the life sciences, but basic concepts like randomisation, blocking and replication can be applied in various fields. Design of experiment is important whenever variation comes into play. Microarray experiments are known to be extremely noisy. Thus it is not astonishing that soon after the technology was established, the discussion about optimal microarray design started (Kerr & Churchill, 2001; Dobbin & Simon, 2002; Speed & Yang, 2002). Some of the specific microarray designs are described in the following section. As they are of limited use if the design objective is the estimation of heterosis, we will first consider some classical optimality criteria. The idea of a criterion that is to be optimized will be adopted and customized to our purpose. We take a closer look at the analysis of the data to determine a suitable optimality criterion. Finally, an optimization algorithm is illustrated that helps us finding a tailor-made design for the detection of genes showing heterotic effects.

The second section is dedicated to the application of these results to a real-life problem: An optimal design for a microarray experiment is searched (Section 2.1.1). We pursue two strategies that allow the detection of differential gene expression between hybrids and their parental inbred lines in maize. Furthermore practical aspects are included: It is demonstrated that results of other experiments may be used, e.g., to check if certain effects should be considered in the design, or to get information about the variance components of the linear model underlying the design. The results of our design investigations may also be found in Keller, Emrich, Höcker, Hochholdinger, and Piepho (2005).

3.1 Design theory

3.1.1 Specific microarray designs

The design of microarray experiments has been the subject of various recent articles. Specific types of designs, such as the common reference design and the loop design, have been proposed (Kerr & Churchill, 2001; Kerr, 2003). As the name suggests, for the common reference design one of the two probes hybridised on one array is a reference sample. This sample is not of primary interest for the experiment. The design is illustrated in Figure 3.1a, where circles represent samples and arrows represent arrays. The sample adjoining to an arrowhead is always labelled with the same colour (e.g., red), the sample adjoining to the contrary end of the array is labelled with the other colour (e.g., green). On each array, only one of the probes contains a treatment that is to be analysed. As the reference sample is always labelled with the same color, the effects of the dye and

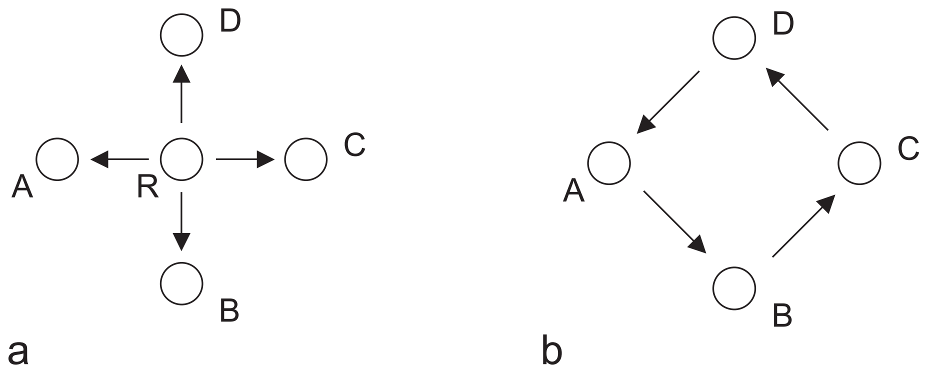


Figure 3.1: *(a) Common Reference Design (b) Loop Design, each with four treatments and four arrays*

treatment are completely confounded. Contrasts of two treatments may be estimated by comparing both probes in relation to the reference sample, e.g., $\tau_A - \tau_B = (\tau_A - \tau_R) - (\tau_B - \tau_R)$, where τ_A and τ_B are the treatment effects and τ_R is the effect of the reference sample. The fundamental handicap of the common reference design is obvious: Half of the probes contain information about a sample that is not of interest. As the hybridisation of microarray experiments is extremely time and cost consuming, this design today is hardly applied.

A popular alternative is the loop design (Figure 3.1b). The arrays are hybridised in a way that each sample is labelled red on one array and green on another array. Each sample is hybridised with two different samples. Contrasts may either be estimated directly from one array, e.g., by $\tau_A - \tau_B$. Others must be estimated indirectly over several arrays, e.g., by $\tau_A - \tau_C = (\tau_A - \tau_B) + (\tau_B - \tau_C)$. This design is more efficient as with the same number of arrays each treatment is replicated. However, if a hybridisation fails (which occurs quite often), the accuracy of estimates decreases substantially.

Modifications of the loop design have been proposed, as the saturated design, where arrays with all treatment combinations exist, or the swapped loop design, where the loop design is hybridised twice with contrary labelling. However, the objectives of these designs differ in one major aspect from the problem we address: In most previous work, solely contrasts between two mRNA populations are considered, whereas for dominance estimation contrasts include more than two genotypes. This procedure complicates the problem since only two of the involved genotypes can be hybridised with one array. This is because for economic reasons most experiments are performed with the fluorochromes Cy3 and Cy5. Classical microarray designs have been applied for problems concerning expression between hybrids and parents, e.g., the loop design applied by Gibson et al. (2004) to assess the degree of additivity in gene expression in *Drosophila melanogaster*. While these designs work, they are usually not optimal with respect to the specific contrasts of interest.

3.1.2 Optimality criteria

Considering the vast literature of specific microarray designs it should not be forgotten that basic principles of design are well-established and can be adopted. For example, a microarray design may be understood as a special case of a row-column design with dyes and arrays corresponding to rows and columns. For this kind of problem optimal designs may be found by numerical search (John & Williams, 1995), e.g., by simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) or tabu search (Glover & Laguna, 1997). These algorithms search the space of possible designs until a near-optimal design is found. In studies where pairs of treatments (genotypes) are compared, one often assumes that all pairwise comparisons are of equal interest, as in Kerr and Churchill (2001). Then criteria such as A-optimality or E-optimality can be applied (X. Yang, Ye, & Hoeschele, 2002; John & Williams, 1995, p. 31).

Design optimality is measured in the estimates' degree of accuracy that will be achieved when analysing the experiment. Usually, this is done by optimizing the 'largeness' of the information matrix $\mathcal{I}$. We consider two criteria based on $\mathcal{I}$ (Pukelsheim, 1993, p. 135).

The D-criterion (determinant criterion) $\phi_D(\mathcal{I})$ is the s-th root of the determinant of $\mathcal{I}$:

$$\phi_D(\mathcal{I}) = (det\mathcal{I})^{1/s},$$

where $\mathcal{I}$ is a $s \times s$ matrix. Maximizing the D-criterion is the same as minimizing the dispersion matrix since $(\det \mathcal{I})^{-1} = \det(\mathcal{I}^{-1})$. The determinant of the inverse information matrix is also called generalized variance. In a linear model setting, the determinant criterion may be visualized by constructing the confidence ellipsoid of the model parameters. The volume of the ellipsoid is inversely proportional to $\phi_D(\mathcal{I})$. Thus, the ellipsoid is smallest when $\phi_D(\mathcal{I})$ is maximized. The popularity of the determinant criterion may at least partly be ascribed to its computational efficiency.

The A-optimality criterion $\phi_A(\mathcal{I})$ is also called average-variance criterion, which already tells a lot about its form:

$$\phi_A(\mathcal{I}) = \frac{1}{\text{tr}(\mathcal{I}^{-1})}.$$

Maximizing $\phi_A(\mathcal{I})$ is synonymous to minimizing the average variance of the parameters to be estimated. The A-criterion is equally simple to compute as the D-criterion because only the diagonal elements of the dispersion matrix need to be computed.

When estimating dominance effects between inbred lines and hybrids, the choice of design is not obvious. Applying classical design theory (Pukelsheim, 1993), one can show whether a design is optimal with respect to the optimality criteria described above. Although these criteria could be applied for our purpose, they are general purpose measures that are more appropriate, e.g., when all pairwise comparisons among treatments are of equal interest. With nuisance parameters in the model (resulting, e.g., from the greenhouse design), these designs would not be optimal for our purpose. Also, optimality of the designs refers to fixed effects models, as usually no information about variance components is given. Luckily we can use previous knowledge about variance components and will therefore consider some effects to be random in later analysis. A more specific optimality criterion is thus preferable (John & Williams, 1995, p. 34). Pearce (1974) and Freeman (1976) propose to minimize the weighted mean of either the efficiency factors of interest or the variance of the contrast of interest. These approaches will be more convenient for our purposes. Due to the relatively large number of factors and factor levels, however, the number of possible designs is very high and checking them all for the best would be computationally very intensive. Landgrebe, Bretz, and Brunner (2006) therefore start with a group of initial designs. These are combined to a set of composite designs, whereof the most efficient one is selected. Another method of reasonably limiting the set of initial designs is by regarding only cyclic designs. Among these optimal designs (M,S)-optimal ones may be readily obtained (John & Williams, 1995). Another possible approach to an optimal design is by determining an upper bound for the average efficiency factor and stop the design search when a design is found that is reasonably close to the optimal design. Such an algorithm is implemented, e.g., in the design generation packages ALPHA+ (Williams & Talbot, 1993) and CycDesigN (Whitaker, Williams, & John, 2002). The design problem we face has some features that make the use of the above mentioned packages difficult or inappropriate: we have several factors whereof only one is of interest, the contrasts of interest imply three levels with unequal weights, and the analysis will be performed by a linear model with fixed and random effects. We therefore decided for a problem-specific optimality criterion and applied two different approaches: one is based on simplification of the design problem and the second uses simulated annealing, a probabilistic optimization algorithm. Like the approaches mentioned in the preceding section, both strategies are trying to find acceptable designs without evaluating every possible design. For the formulation of the optimality criterion adapted to our problem it is helpful to consider the linear model analysis of the data which will be gained by the experiment.

3.1.3 Linear models and linear mixed models

According to Mead (1988) the design of an experiment should be closely linked to its analysis. Similar to the identification of differentially expressed genes (Dudoit, Yang, Speed, & Callow, 2002) the determination of dominant genes is done by a linear model.

Microarray data are frequently analysed by a linear model, which is described by

$$y = X\beta + e,$$

where y is a vector of observations with n elements, X is the design matrix, which can either contain continuous or categorical variables, β is the parameter vector of fixed effects, and e is a vector of random error with $\mathrm{E}[e] = 0$ and $\mathrm{Var}[e] = \Sigma$, where $\Sigma = \sigma^2\mathrm{I}$, i.e., the elements of the error vector are i.i.d.

Especially in an experiment with more than two genotypes it is helpful to regard the array effect as random, resulting in a linear mixed model, which is described by

$$y = X\beta + Zu + e,$$

where Z denotes a known design matrix and u stands for the vector of random effects (Searle, Casella, & McCulloch, 1992, p. 233). It is specified by $\mathrm{E}[u] = 0$ and $\mathrm{Var}[u] = D$. Therefore, y is distributed with mean $X\beta$ and variance $\mathrm{Var}[y] = V = ZDZ' + \Sigma$, where Σ now may be an arbitrary variance-covariance matrix.

The random array effect is useful for the following reason: in an experiment where more than two genotypes are to be compared, arrays may be regarded as incomplete blocks. Contrary to a fixed effects model where contrasts between genotypes are estimated using solely information on comparisons within a block, with mixed models and incomplete blocks the recovery of inter-block information is possible (John and Williams (1995, p. 27) and Cochran and Cox (1957, p. 382)). When the variability between blocks (or arrays) is low, including the inter-block analysis may achieve a substantial gain in accuracy of estimates. By contrast, when the block variance is high, contrast estimates will largely result from the intra-block analysis.

Suppose we have r random effects which are mutually independent, i.e., $D_{ii'} = 0$ for $i \neq i'$ where $D_{ii'}$ is the covariance of the random effects u_i and u_i'. Then D is a diagonal matrix $\{_dD_i\}_{i=1}^r$ and the variance is $V = \sum_{i=1}^r Z_iD_iZ_i' + \Sigma$.

As the observations are assumed to be normally distributed, the log-likelihood is characterized by

$$l = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)' V^{-1}(y - X\beta) - \frac{N}{2}\log(2\pi). \tag{3.1}$$

If V is known, the fixed effects may be estimated by differentiating the log-likelihood with respect to β. As the derivative of a quadratic form $x'Ax$ with respect to x is $2Ax$ for a symmetric matrix we have

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= -\frac{1}{2}(-X') \cdot 2V^{-1}(y - X\beta) \\ &= X'V^{-1}(y - X\beta).\end{aligned}$$

Equating the derivation to zero gives the ML equation

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{3.2}$$

However, in practice V usually is unknown. Thus the log-likelihood is not only differentiated with respect to β but also with respect to the variance components σ_i^2 in V. Using $\frac{\partial}{\partial\sigma_i^2}V = Z_iZ_i'$, $\frac{\partial}{\partial\sigma_i^2}\log|V| = \text{tr}(V^{-1}\frac{\partial V}{\partial\sigma_i^2})$, and $\frac{\partial}{\partial\sigma_i^2}V^{-1} = -V^{-1}\frac{\partial V}{\partial\sigma_i^2}V^{-1}$ we get:

$$\frac{\partial l}{\partial \sigma_i^2} = -\frac{1}{2}\text{tr}\left(V^{-1}Z_iZ_i'\right) + \frac{1}{2}(y - X\beta)'V^{-1}Z_iZ_i'V^{-1}(y - X\beta).$$

This expression is equated to zero for each variance component $\sigma_i^2, i = 1, ..., r$, giving

$$\text{tr}(V^{-1}Z_iZ_i') = (y - X\beta)'V^{-1}Z_iZ_i'V^{-1}(y - X\beta). \tag{3.3}$$

Equation (3.2) and (3.3) usually must be solved numerically to obtain a solution for β and σ_i^2. We can reduce the problem and write it in a simpler form. With (3.2) we have

$$V^{-1}(y - X\beta) = V^{-1}(y - X(X'V^{-1}X)^{-}X'V^{-1}y), \tag{3.4}$$

where $(X'V^{-1}X)^{-}$ is a generalized inverse and (3.4) is invariant with respect to the generalized inverse. We now define

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1} \tag{3.5}$$

and after some arithmetic we get

$$\text{tr}(V^{-1}Z_iZ_i') = \quad y'PZ_iZ_i'Py \tag{3.6}$$

(Searle et al., 1992, p. 236). This equation has to be solved for β and σ_i^2 numerically, giving the ML estimates $\hat{\beta}$ and $\hat{V}$.

A drawback of Maximum Likelihood estimates is that the loss of degrees of freedom caused by the estimation of fixed effects is not taken into account and hence variance components are underestimated (Searle et al., 1992, p. 249). This problem may be addressed by the REML (restricted maximum likelihood) approach. With REML-estimation, random effects are estimated by maximizing the likelihood of linear contrasts of elements of y. The linear combinations are chosen in a way that fixed effects are eliminated, i.e. $\mathrm{E}(k'y) = 0$ where k' denotes a contrast vector. There are $n - \text{rank}(X)$ linearly independent vectors with this property, which are all used for the estimation of variance components to yield optimal results. The matrix of contrast vectors is written $K = [k_1k_2...k_{n-\text{rank}(X)}]$. As $K'y \sim N(0, K'VK)$, the REML-likelihood is

$$l_R = -\frac{1}{2}\log|K'VK| - \frac{1}{2}(K'y)'\,|K'VK|^{-1}\,K'y - \frac{n - \text{rank}(X)}{2}\log(2\pi).$$

This is known as the marginal likelihood and is not dependent on the fixed effects β. The REML-estimates are obtained by maximizing l_R. In accordance to (3.1) they may be derived by replacing y by $K'y$, Z by $K'Z$, X by

$K'X = 0$ and V by $K'VK$ in (3.6) leading to

$$tr((K'VK)^{-1}K'Z_iZ_i'K) =$$
$$y'K(K'VK)^{-1}K'Z_iZ_i'K(K'VK)^{-1}K'y$$

for each $i = 1, ..., r$. According to Khatri (1966),

$$V^{-1} - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1} = K(K'VK)^{-1}K'$$

and thus

$$\text{tr}(PZ_iZ_i') = y'PZ_iZ_i'Py$$

for each $i = 1, ..., r$. To gain estimates for the variance components, the REML-equations are to be solved numerically, leading to $\hat{V}_R$, the REML-estimate of V. In SAS/ Proc Mixed this optimization is done by a ridge-stabilized Newton-Raphson algorithm.
The REML-approach does not include a method to estimate fixed effects. Usually the ML equation for the fixed effect is used with $\hat{V}_R$ instead of the ML estimate $\hat{V}$. Therefore

$$(X\hat{\beta})_R = X(X'\hat{V}_R^{-1}X)^{-}X'\hat{V}_R^{-1}y$$

may be used to estimate $X\beta$ and

$$\text{Var}(X\hat{\beta})_R = X(X'\hat{V}_R^{-1}X)^{-}X'$$

is the asymptotic variance-covariance matrix.

The REML-approach is sometimes preferred over the ML approach, because of the above mentioned property to consider the degrees of freedom lost by estimating the fixed effects. This leads to estimates of V that are less biased compared to ML-estimates. Furthermore, as the REML-likelihood does not depend on β, the values of the fixed effects do not influence the estimates for the variance components. A third merit of the REML-estimators is that REML-estimators seem to be less sensitive to outliers than ML estimators (McCulloch & Searle, 2001, p. 177-178).

Suppose the variance of a contrast $l'\hat{\beta}$ is to be estimated. As $\text{Var}(X\hat{\beta}) = X(X'V^{-1}X)^{-}X'$ where $\hat{\beta}$ is the ML estimator of β, $\text{Var}(l'\hat{\beta}) = l'(X'V^{-1}X)^{-}l$. For unknown V, V may be replaced by $\hat{V}$.

If we choose $l'\beta$ as the heterosis contrast of (1.1), then this is exactly what we want to estimate by a suitably chosen design. We take the standard error of the heterosis contrast as optimality criterion for the design:

$$\text{SE}(l'\hat{\beta}) \sim \sqrt{\left(l'(X'\hat{V}^{-1}X)^{-}l\right)}, \tag{3.7}$$

A design is considered optimal, when the standard error of contrast (1.1) is lowest.

Proceeding with the analysis, one certainly wishes to make inference about the heterosis contrast by testing the hypothesis $l'\beta = 0$ against the alternative $l'\beta \neq 0$ with the test statistic

$$t = \frac{l'\hat{\beta}}{\sqrt{l'(X'\hat{V}^{-1}X)^{-}l}}.$$

According to McLean and Sanders (1988) t is approximately t-distributed. The degrees of freedom may be approximated by the methods of Satterthwaite or of Kenward-Roger. The Satterthwaite-option implemented in SAS/Proc Mixed is a generalization of the methods described in Giesbrecht and Burns (1985), McLean and Sanders (1988) and Fai and Cornelius (1996). The Satterthwaite method is sometimes unsatisfactory as the dispersion matrix of estimated fixed effects is underestimated (Kackar & Harville, 1984). The method of Kenward and Roger (Kenward & Roger,

1997) improves the Satterthwaite method with a correction of the dispersion matrix, which is especially valuable for small samples. Spilke, Hu, and Piepho (2005) found through a simulation study that the underestimation of the dispersion matrix is substantially reduced when applying this correction. Furthermore, the method of Kenward and Roger (1997) showed the best control of the Type I error compared to other approximations of the degrees of freedom, being competitive in terms of power. It is therefore recommended to use the approximation of Kenward and Roger.

Besides the ML- and REML-approach, the analysis of variance (ANOVA) method is a third method of analysing linear models. The total sum of squares of the data is split into sum of squares for the factors and the residual sum of squares. Expected sum of squares are equated to observed sum of squares and the resulting system of equations is solved for the variance components. In Chapter 5 the sum of squares concept is used for an empirical Bayes approach to variance component estimation.

3.1.4 Simulated annealing

As we defined an optimality criterion, we need a strategy to find the design with the best value of the criterion. One could perform a complete search of all possible designs. With larger problems, however, this is not feasible. In these cases numerical search methods such as the simulated annealing algorithm (SA) may be applied (Kirkpatrick et al., 1983). SA is an algorithm for the global optimization of a function. It originates from metallurgy: If a piece of metal is annealed slowly, its atoms arrange in a way that the emerging crystal structure has minimum energy. If the metal is cooled down too fast, the atoms do not have enough time to arrange in a low-energy crystal lattice. In this case the atoms are stuck in a local optimum. If the atoms are arranged in a crystal structure with lowest energy, the metal piece is very stable.

How may this procedure be carried forward to function optimization? Firstly, the objective function to be optimized must be chosen. In case of design optimality, the optimality criterion of the design is regarded as objective function. We use the design to estimate dominance effects with microarray data and assume the signal intensities to be influenced by certain effects via a mixed model. Therefore, it is reasonable to use the standard error of dominance contrasts as objective function.

We evaluate the optimality criterion for a random start design. Then a random change in the design matrix is performed, i.e., the effect to be altered as well as the new level of this effect is chosen randomly. If the design has improved or, in other words, the optimality criterion has decreased, the new design is accepted and another random change is performed. Otherwise, if the new design is worse, it is not discarded in every case, but accepted with a certain probability p. In the next step, either the new design, or, in case of rejection, the old design is altered, and so on. Accepting a design in some cases even if it is worse allows moving away from a local minimum. The acceptance probability is dependent on the difference between the optimality criteria of the design before and after the variation:

$$p = \exp\{-[f(D_{new}) - f(D)]/T\}, \tag{3.8}$$

where T is a variable analogous to the temperature in the annealing process explained below and $f(D)$ and $f(D_{new})$ are the values of the objective function of the present design and the altered design, respectively. If the worsening of the design is serious, the probability of rejection of the new design is higher than with an only slightly inferior design. The temperature T decreases in successive iterations. Thus, with respect to the temperature, the acceptance probability is close to one at the beginning of the process when the temperature is high. Hence, the new design will very often be accepted. When the temperature approaches zero, the exponent of (3.8) will draw near minus infinity and the acceptance of an inferior design is unlikely. When the design has not changed for a certain number of tries, the algorithm stops. Details of the simulated annealing algorithm may be found in Kirkpatrick et al. (1983) and an application to design search including the algorithm in pseudocode is given in Angelis, Bora-Senta, and Moyssiadis (2001). Independently of our work (Keller et al., 2005), simulated annealing has been applied to microarray design problems by Wit, Nobile, and Khanin (2005).

3.2 Application

Material and methods

This section will illustrate how a microarray design may be developed for a real-life problem. The task was to determine an optimal design for a microarray experiment to estimate differential gene expression between hybrids and their parental inbred lines of maize, as described in Section 2.1.1. The precise definition of the objectives of the study leads us to a definition of design optimality within the given context. The main steps of the planned experiment are described in detail to account for all effects that might influence hybridisation signals. These effects are included in the model used for the design search. To determine the significance of these effects, we used data from a pre-experiment (Section 2.1.3). Information on the variance components was derived by analysing a microarray experiment that had previously been conducted in the same laboratory (Section 2.1.2). Finally, we explain two methods to find designs with the defined optimality properties. Without a doubt, other microarray studies are carried out in a different manner, and some of the effects we account for will not emerge. It should be stressed, however, that with the mixed-model approach other effects can easily be included in the model.

The design problem has two characteristics: Contrary to the main application of microarray analysis where two genotypes or treatments of equal interest are compared, here the contrasts of interest contain more than two genotypes. This makes necessary a newly defined optimality criterion and a tailor-made strategy to search the design space. We employ the mean standard error of dominance contrasts (3.7), calculated by the restricted maximum likelihood method, as optimality criterion. Two approaches were used to find an optimal design: the first one simplifies the problem by dividing it into several subproblems, whereas the second is more sophisticated and uses a simulated annealing algorithm. The second characteristic is that the procedure may be customized to other microarray experiments where different effects may influence hybridisation signals. A mixed model was used to include all important effects. Impacts during growth of plant material were taken into consideration as well as those occurring during hybridisation. By means of a preliminary experiment it was decided which effects are to be included in the model and data from another microarray experiment were used to estimate variance components.

Planned experiment

To account for all of the effects that might influence cDNA samples, a knowledge of their origin is of utmost significance. In the planned experiment, 20 maize seeds germinate together in a role of filter paper (see 2.1.1). Several of these filter paper roles with seedlings are placed in a water filled beaker. In order to harvest all seedlings at approximately the same time of day and thus avoid circadian effects, the number of paper rolls in one beaker is limited to 16. After 84 hours, mRNA is extracted from the roots of the germinated seedlings and transcribed into cDNA. The cDNA is hybridised onto the microarrays and the array is scanned to get information about the signal intensity. We assume that there is a roughly log-linear relationship between the amount of expression product and the signal detected by the scanner. The experiment was planned for a total of 72 microarray chips. Effects that occur during this procedure and which might influence hybridisation signals are included in the following linear mixed model:

$$y_{ijkl} = \mu + g_i + d_j + (gd)_{ij} + b_k + c_l + e_{ijkl}. \tag{3.9}$$

Here, for $i = 1, ..., n_i$, $j = 1, 2$, $k = 1, ..., n_k$ and $l = 1, ..., n_l$, y_{ijkl} is the log signal intensity for genotype i on array l, marked with dye j. Plant material for this sample was cultivated in beaker k. Further definitions are:

μ, the overall mean;
g_i, the fixed effect of genotype i;
d_j, the fixed effect of dye j;
$(gd)_{ij}$, the interaction between genotype i and dye j;
b_k, the fixed effect of beaker k;

c_l, the random effect of array l;
e_{ijkl}, the random residual error associated with y_{ijkl};
n_i, n_k and n_l, the numbers of levels of the corresponding effect.

An effect of filter paper can be taken into account as well, but we found in the root length experiment described in section 2.1.3 that this effect is not significant. When the array effect is treated as random, the recovery of inter-array information becomes possible. This may result in more accurate estimates of contrasts between inbred lines and hybrids, depending on the magnitude of the variance component involved and the associated degrees of freedom. Contrary to the present study, the recovery of inter-array information (analogous to inter-block information in incomplete block designs; see John and Williams (1995, p. 27)) is not an issue in experiments studying only two treatments, where arrays constitute complete blocks.

In the preceding sections we developed the standard error of the dominance contrast (3.7) as a suitable optimality criterion for our design. As we have more than one hybrid, we computed the mean of the standard errors over the 12 hybrids for each potential design. The array variance/residual variance-ratio was provided by an earlier microarray experiment, which will be reported in a following section (experiment 3). Here it is not possible to perform the experiment with biological replicates in the sense that each sample consists of RNA of one single maize plant. As field design is not known, we only have RNA from a pool of plants with a certain genotype. However, biological replicates allow the investigator to make an inference on the population from which the replicates derive and should be used whenever possible. One would then include a replicate effect in the model to account for variance between individual biological replicates.

Pre-Experiment for significance testing of possible effects

A pre-experiment was performed to assay the influence of filter paper and beaker, which may arise during the germination of the seedlings (Section 2.1.3). Effects of filter paper and beaker were incorporated in a mixed model. We assume that results from the pre-experiment, which are based on phenotypic data, also apply to the gene expression level. As reported in the results section, the pre-experiment revealed no significant effect of the filter paper, whereas the influence of the beakers was confirmed in the pre-experiment. Therefore, regarding the experimental design for the planned microarray experiment, we did not account for a filter paper effect.

Estimating variance components from an earlier microarray experiment

To collect information about the variances between and within arrays, we analysed data from a microarray experiment where two maize genotypes (wild type and the mutant rtcs (Hetz, Hochholdinger, Schwall, & Feix, 1996)) were examined for differentially expressed genes (see 2.1.2). The experiment was carried out according to the same protocol and in the same laboratory as will be Experiment 2.1.1. Then, analysis was performed for every gene according to a mixed model including effects for genotype, dye, array and genotype-by-dye interaction, the array being the only random effect. We thus obtained estimates for array variance as well as for residual (within-array) variance. Medians of both estimates were used for later design considerations where we used the so determined ratio of variance components.

Finding optimal designs

In Experiment 2.1.3 we showed that the filter papers, in which the maize seeds were germinated, have no major influence on root length. Thus, it is reasonable to germinate only one genotype per filter paper, instead of using filter paper as a blocking variable. This simplifies the experimental design considerably. With the filter paper having no significant influence, the design problem is the following: How should genotypes be allocated to the cDNA-samples, how should the two dyes be allocated to genotypes, and how should the filter papers be assigned to the beakers to achieve low standard errors for the contrasts?

We addressed the design problem in two steps. As we have six pairs of hybrids and reciprocal hybrids, we formed six groups containing hybrid, reciprocal hybrid and parents. For example, the first group would comprise A, B, AB and BA, the second A, C, AC and CA and so on. The groups were denoted as ′A-B′, ′A-C′, etc. For estimating the dominance contrasts of a certain hybrid and its reciprocal one group is sufficient. For example, to estimate $\delta(AB)$ and $\delta(BA)$, only the first group is necessary. Also, each hybridisation of two genotypes can be uniquely allocated to a certain group, e.g., an array with genotypes A and AC is said to be in the second group.

With a total of 72 arrays we have 12 arrays available for every hybrid-reciprocal group. To have similar experimental conditions for all samples it would be preferable to germinate all seeds in the same beaker. However, for lack of space the number of filter papers per beaker is limited and two beakers per group are needed.

To find a good design one approach is to search for an optimal design for one group, i.e. indicate the optimal number of replicates of the six combinations of the four genotypes (A-AB, B-AB, A-BA, B-BA, AB-BA, A-B) as well as the optimal allocation to beakers and dyes. To reduce the number of possibilities we imposed a restriction: One half of the replicates with a certain genotype pair, e.g., A-AB, should be grown in each beaker and, accordingly, with half of the replicates of a certain genotype pair the dyes should be swapped. This restriction excludes highly unbalanced designs, which are expected to be inferior regarding the optimality criterion, and the number of possible designs to evaluate is computationally feasible. We generated and evaluated all possible designs in this restricted set and picked the best. Then this optimal design was adapted to the other groups by inserting the appropriate genotype identifiers. Finally, all generated design matrices were composed to a matrix including all genotypes. The resulting design will further be denoted as 'compound design'.

The compound design neglects the fact that a parent does not only occur in one group, but in three. Combining information of groups will increase the information about the parents and therefore the dominance contrast. Hence, the compound design might not be optimal for the whole problem. As computing and evaluating of all possible full designs (72 microarrays, four effects) is very time-consuming, we performed the search with a simulated annealing algorithm (Section 3.1.4). Providing a start design, the algorithm performs a random change in the design matrix, i.e. an array and an effect (of either genotype, dye or beaker) to be changed is randomly chosen. If the beaker effect is chosen, then a second array currently allocated to the other beaker is picked and swapped with the first array. This ensures the same number of filter papers in both beakers. The idea of forming groups of hybrid, reciprocal and parents is kept in the sense that, when altering the genotypes hybridised to an array, the 'new' genotypes must be of the same group as the former genotypes. But, unlike the first approach, optimisation is done for all genotypes simultaneously. The 'start temperature' was chosen $T_0 = 1$ and annealing was conducted by multiplying the current temperature with 0.95 in each iteration.

To analyse the usefulness of the chosen optimality criterion we compared the design satisfying this criterion with an A- and D-optimal design for genotype effects. We included the three effects of genotype, array and dye and searched for the optimal design for one group (i.e. for 12 arrays) in each case. Both the cases of fixed and random error effects were evaluated. Furthermore we varied model (2.1.1) underlying both SA- and compound design and considered the consequences for the complete design. We assumed the array effect to be fixed or random with different variance components and omitted the beaker effect.

Analysis of the experiment

Our project partners (Frank Hochholdinger, University of Tübingen) decided to use the compound design for their microarray experiment. Each microarray slide was scanned six times to obtain optimum information about weakly expressed spots as well as about spots with high signal values. To combine data from different slides a nonlinear regression model was applied Piepho, Keller, Höcker, and Hochholdinger (2006) (Section 6.3). After normalization of the data (Section 2.3), the analysis was performed for each spot according to model (3.9). We computed t-tests for the hypotheses H_0: 'Gene expression differs between hybrid and parental mean' and the alternative H_A: 'Gene expression does not differ significantly between hybrid and parental mean'. The resulting p-values were adjusted for multiplicity with the false discovery rate (Benjamini &

Hochberg, 1995).

3.2.1 Results

Analysis of pre-experiment 2.1.3 data showed significance of the fixed effect for beaker (p-value 0.0218). Evaluating the random effect for filter paper we found no significance. This had the important consequence that we could choose the simplest way of cultivating plants for one sample, i.e., cultivate plants on the same piece of filter paper. If the filter paper effect had been significant, it would have been worthwhile to use the filter paper as a blocking variable.

We obtained estimates for array and residual variance. After computing medians for both variance components, we took the relation array variance $\approx 0.48\times$residual variance for further calculations.

The results of these preliminary analyses were used to parameterise the model with which the design was optimised. The first solution is a design generated by optimising the sub-design for each group and then piecing together sub-designs. Therefore, designs for every group have the same number of replicates of hybrid-parent, reciprocal-parent, hybrid-reciprocal and parent-parent hybridisations. The second solution, optimised for the full design, was obtained by an SA-algorithm. Again, the design has the same number of replicates for every group, although here it is not pre-determined as in the first solution.

We first note that with both approaches, the selected design has no parent-parent arrays (Figure 3.2). The reason is that this pair does not provide any information on the dominance contrast. Yet, the parent-parent contrast can be estimated with good accuracy, because the designs provide

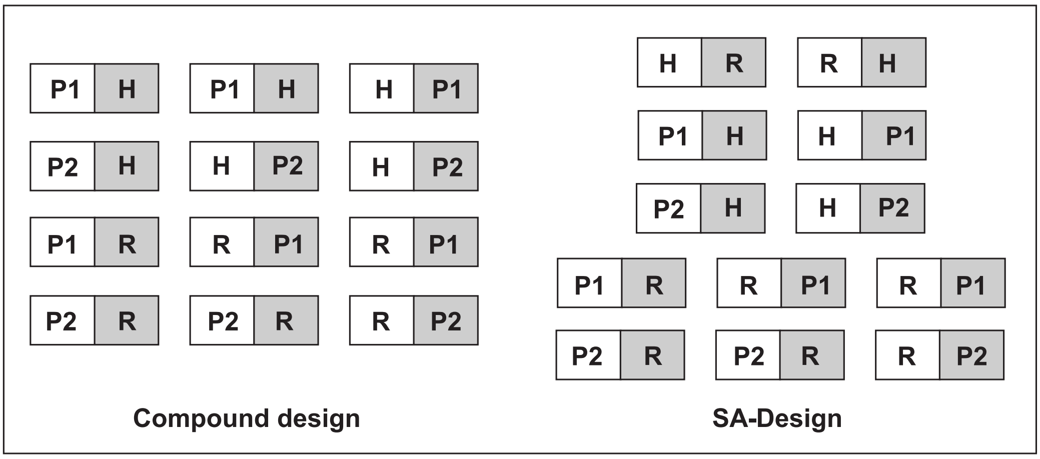


Figure 3.2: *Diagram indicating hybridisations and labelling directions for the compound-approach and SA-approach for one group (white $\hat{=}$Cy3, grey $\hat{=}$ Cy5)*

many indirect comparisons among the parents via the hybrids. For example the contrast A-B can be estimated from the difference of the contrasts A-AB and B-AB or from contrasts A-AC, C-AC, B-BC, and C-BC. Similarly, Piepho (2005) found that when estimating heterosis, parent-parent pairs or hybrid-reciprocal pairs should be used sparely or not at all to obtain accurate heterosis estimates.

It is striking that with the compound design we do not have any hybrid-reciprocal hybridisations while in the SA-design there are two per group. The explanation is that in the SA-approach we also exploit information about the parents available from other groups, where the same parents occur. Thus, fewer parents need to be hybridised and hybrids are used instead. As a certain hybrid only appears in one group, it makes sense to increase the number of hybrid hybridisations. A closer look at one group of the SA-design (Table 3.1) reveals that there is a dye swap across beakers except in the third row where the parent changes. Due to this change the number of both parents is balanced.

Table 3.1: *Allocation of genotypes*[§]*, beakers and dyes exemplified for one group of the SA-design.*

Beaker 1		Beaker 2	
Cy3	Cy5	Cy3	Cy5
P1	H	H	P1
P2	H	H	P2
H	P1	P2	H
R	P1	P1	R
R	P2	P2	R
H	R	R	H

[§] P1, P2: parents; H, R: hybrid and reciprocal cross.

We also see that in the SA-approach, we have unequal numbers of replicates for hybrid-parent and reciprocal-parent hybridisations. Consequently, with this design the dominance contrast for a hybrid cannot be estimated with the same accuracy as the dominance contrast for the reciprocal. Of course, hybrid and reciprocal hybrid are interchangeable. Therefore, it is possible to estimate the favoured dominance contrast with greater accuracy. Parental contrasts are estimated with varying accuracy depending on the genotypes. The variations may be caused by different dye- and beaker-allocations. These allocations do not show any systematic pattern as can be seen from the allocation of genotypes to the arrays. Standard errors for hybrid-reciprocal contrasts are the same for every group, as we always have within each group one hybrid hybridised six times and one hybridised four times.

In Tables 3.2 standard errors of different contrasts are given for the two designs. As an unequal number of hybrids and reciprocal hybrids was hybridized in the SA-design, the standard errors of the dominance contrast for hybrid and reciprocal hybrid are different. The optimality criterion was calculated as the mean of standard errors of the dominance contrast for hybrids and reciprocal hybrids. It is not astonishing that the value of the optimality criterion is worse with the compound-design approach, because the design was optimised for only one group and not the problem as a whole. As the optimality is worse, standard errors for dominance contrasts are higher than the mean of standard errors for the SA-approach. Only parental contrasts are estimated better in the first approach, which seems plausible as parents are hybridised more often.

Table 3.2: *Effectiveness of the two approaches (complete design)*

	Standard errors	
	Compound design	SA-design
Dominance contrast		
Mean (optimality criterion)	0.5268	0.5256
Range	0.5268	0.4979 or 0.5534
Parental contrasts (range)	0.4802	between 0.5308 and 0.5814
Hybrid-reciprocal contrasts	0.6658	0.5948

The increase in accuracy of estimation when joining information of several groups can be seen when comparing standard errors of a reduced design containing genotypes of only one group with standard errors of the complete SA-solution (Table 3.3). By combining all groups the gain in accuracy of estimation for the dominance contrasts is rather small. Especially the parental contrasts are estimated more accurately when taking the complete design, as we have altogether three groups which provide information about a parent. The accuracy of contrasts between hybrid and reciprocal differs only slightly between designs because no other hybridisations are of interest than those with the parents of the according group.

Table 3.3: *Comparison of one-group- and complete SA-design*

	Standard errors	
	One group of SA-design	Complete SA-design
Dominance contrast	0.5027 or 0.5555	0.4979 or 0.5534
Parental contrast	0.7478	0.5469
Hybrid-reciprocal contrast	0.5952	0.5948

The increase in accuracy achieved with the simulated annealing (SA) approach is relatively small (Table 3.2). Also, not all contrasts are estimated with the same accuracy. Therefore, the gain from using the SA-algorithm was not dramatic for this experiment. Generally, the gain from the SA-method strongly depends on the factors and their levels included in the model and can hardly be evaluated in advance.

For the evaluation of our optimality criterion we developed an A- and D-optimal design, which has two replicates of each genotype-combination (A-B, A-AB, A-BA, B-AB, B-BA, AB-BA) with the dyes swapped. The design optimal for the heterosis contrasts contains two additional hybrid-parent-replicates instead of the parent-parent replicates. Standard errors for the dominance contrast are 0.5276 (heterosis-optimal) and 0.5466 (A-/D- optimal). This means the variance of the heterosis-optimal design is about 93% of the variance of the A-/D-optimal design. Taking the chip effect as fixed the design optimised for heterosis performs even better compared to the A-/D- optimal design: its variance then is only 87% of the A-/D-optimal design for the heterosis contrast.

Considering the complete design assuming the array effect as fixed changes the compound design (Figure 3.3) but not the design obtained by simulated annealing. A fixed array effect corresponds to a random array effect with infinite variance. Therefore, if the array variance is high compared to the residual variance this makes a difference only for the compound design but not the SA-design. The extreme case of fixed chip effects suggests that with other variance ratios a change in the optimal design is more likely with the compound design than with the SA design. If the beaker effect is omitted, the compound design is not affected, but the SA-approach results in an increased number of hybrid-reciprocal arrays at the cost of hybrid-parent arrays (Figure 3.3). It thus seems justified to account for this effect.

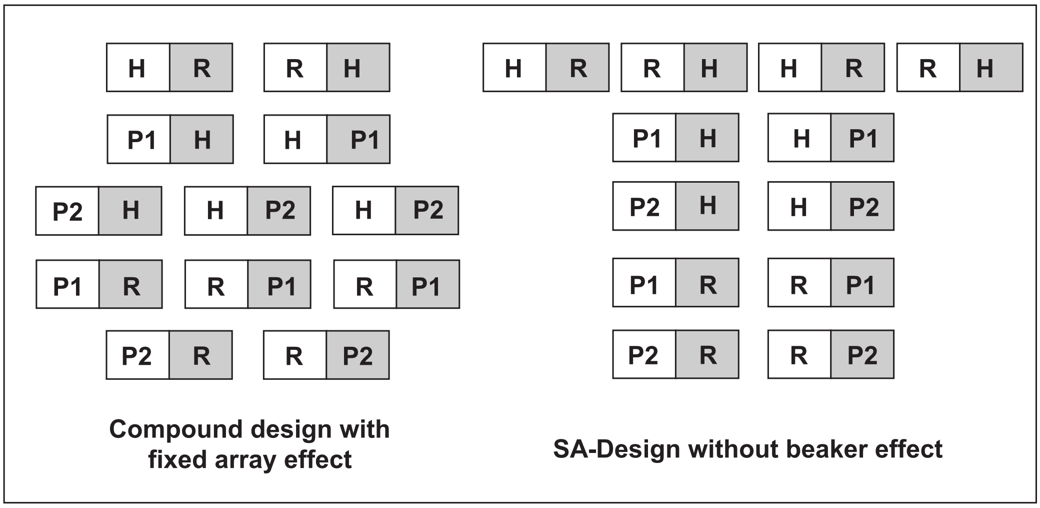


Figure 3.3: *Diagram indicating hybridisations for variations of model (3.9) (white$\hat{=}$Cy3, grey$\hat{=}$Cy5)*

The microarray experiment was performed according to the compound design, whereof a simplified version is displayed in Figure 3.2. Analysis of the data revealed that the hybrids differed largely in the number

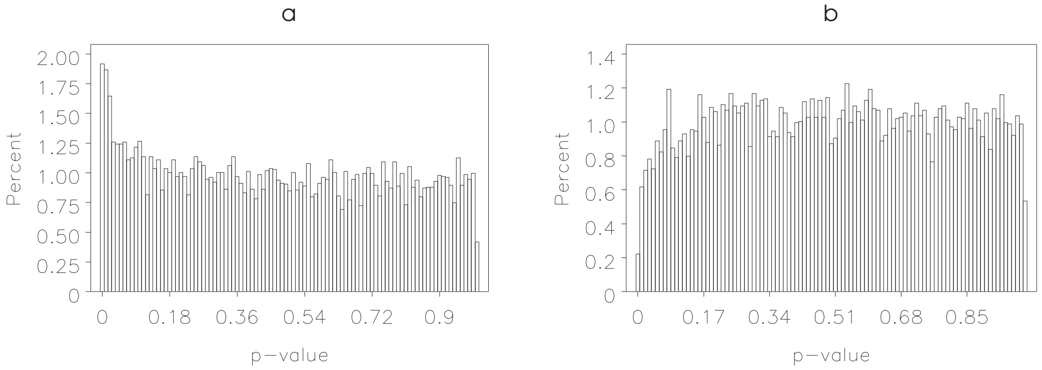


Figure 3.4: *Histogram of p-values for (a) maize hybrid UH250xUH002 and (b) maize hybrid UH250x301*

of genes that show a significant effect for dominance. In seven hybrids no gene with significant dominance contrast (after fdr-adjustment) was found. Among these hybrids there are four intra pool hybrids, i.e., the parental lines are genetically similar and thus a dominance effect is less likely to occur. All of the five hybrids where differences between hybrid and parental mean could be determined are inter-pool hybrids, meaning that the genetic diversity between the parental lines is higher than with intra-pool hybrids. Histograms of unadjusted p-values for two hybrids are shown in Figure 3.4. Hybrid UH250xUH002 is an inter-pool hybrid where—after fdr-adjustment—24 genes had a significant dominance contrast. Hybrid UH250xUH301 is an intra-pool hybrid where no significant dominance contrast was found.

Only for hybrid UH250xUH002 a deviation from the uniform distribution can be seen, resulting in differential contrasts. In Table 3.4 genes with a significant difference in gene expression between UH250xUH002 and the parental mean are indicated together with the estimate of the dominance contrast and the fdr-adjusted p-values.

Table 3.4: *Genes with a significant dominance contrast for hybrid UH250xUH002*

Clone ID	Dominance estimate	p-value
605012F10.x1	1.240	0.017
MEST36-E03	1.050	0.014
606013F12.x2	0.898	0.017
605014A07.x1	0.860	0.026
MEST36-E07	0.842	0.026
603019B04.x1	0.831	0.025
606014H01.x1	0.821	0.030
614095B03.x1	0.774	0.014
605012A06.x1	0.771	0.030
MEST34-H12	0.760	0.018
614018D07.y1	0.757	0.025
606005C03.x1	0.757	0.025
606013G02.x2	0.731	0.017
MEST11-A10	0.731	0.029
605002F03.x1	0.702	0.018
MEST9-F02	0.686	0.043
486066A07.x1	0.658	0.025
606013E04.x2	0.622	0.029
707081C10.x1	0.594	0.022
MEST113-E01	-0.506	0.025
MEST66-A07	-0.640	0.025
606014C11.x1	-0.647	0.048
MEST67-E11	-0.838	0.019
603040C09.x1	-0.907	0.025

Höcker et al. (2007) classified genes with a significant dominance contrast according to their function and found that differentially expressed genes fell in all functional categories. They therefore suppose that not a specific function is required during heterosis manifestation in maize primary roots but rather the interplay of genes related to diverse functions.

3.2.2 Discussion

We sought for a microarray design with minimum standard errors for the desired contrasts. As a first method a solution for a simplified version of the problem was computed. A simulated annealing algorithm was used for optimisation in the second method and a design adapted to the specific problem was provided.

As optimality criterion the mean standard errors of all dominance contrasts was chosen. Other criteria would be possible, according to research objectives. John and Williams (1995, p. 34) propose to choose a criterion weighting the contrasts according to their importance. In our case, zero weight was given to all contrasts except the dominance contrasts, because this conformed to the main objective of the planned experiment. Other weighting schemes comprising standard errors for other contrasts, for example parental contrasts, are imaginable. For example, in order to study the dominance and the over-dominance hypotheses, it is useful to consider the comparison of a hybrid with one of its parents. These contrasts were not of primary interest for the planned experiment since the main objective was to identify genes showing dominance effects.

The optimality criterion evaluated during the design search is an approximation to the mean standard errors of dominance contrasts, as the true variance components in the model are unknown. According to

Kackar and Harville (1984) the variance of a linear contrast of β may be approximated by

$$\text{Var}(l'\hat{\beta}) \doteq l'(X'V^{-1}X)^{-}l + tr\left[C\{{}_m h' Z_i Z_i' P Z_j Z_j' h\}_{i,j=0}^{r}\right], \tag{3.10}$$

with $h' = l'(X'V^{-1}X)^{-}X'V^{-1}$, P as in (3.5) and $C = \{{}_m c_{ij}\}_{i=0,j=0}^{r\ \ r}$ is the asymptotic variance-covariance matrix of the vector of estimated variance components. This approximation accounts for uncertainty in the variance estimates. Instead of (3.10), however, we used

$$\text{Var}(l'\hat{\beta}) \doteq l'(X'V^{-1}X)^{-}l, \tag{3.11}$$

with V replaced by $\hat{V}_R$, the REML-estimate of V, as described in Section 3.1.3. The reason for approximating (3.10) by (3.11) is that this expression is computed directly by SAS/ Proc Mixed.

The discussion shows that choice of an optimal design depends on a number of factors. In addition, the common optimality criteria (D- and A-optimality, average pairwise variance) are not generally helpful. Thus, standard packages for experimental design do not usually give the most useful answer, and a tailor-made approach is needed. Further details regarding this aspect can be found in Pearce (1974) and Freeman (1976).

With the analysis of a pre-experiment as well as a further microarray experiment, we gained knowledge about the significance and magnitude of error effects. Because a significant effect of filter paper could not be proved for phenotypic data, this effect was neglected. Yet it is not clear if this is satisfactory proof that this effect does not show up in mRNA. If so, the filter paper effect will be confounded with the residual intra-array variance and then will increase the error term. Analysis of microarray data showed that array variance is about half of the residual variance. This, however, is an estimate based on another experiment and in the planned experiment the variance ratio may possibly change.

In this study some basic principles, which can generally be used when designing microarray experiments, were applied. First of all, a mixed model underlies all design considerations. Effects for array, dye, and genotype will probably be incorporated in every microarray design. Depending on the way in which plant material is obtained, the inclusion of other effects will be necessary. If one is doubtful which of them are significant, a separate experiment can be performed to check these factors. If information about variance components of the random effects is available from other sources, this can be utilised. Then, after defining an appropriate optimality criterion, the search for the optimal design can be carried out. One approach is to simplify the design problem and choose the best among all designs that satisfy some reasonable restrictions. This simple strategy provides fairly good results compared to a more complex design solution.

This work is the outcome of collaborative efforts within a research network 'Heterosis in Plants' addressing the microarray analysis of young seedling roots in maize. Naturally, other research groups will face different design problems, mainly in the early stages of their projects (e.g., during cultivation of plant material used for hybridisation), but some of the concepts elaborated here still hold, and, with some modifications, results can be applied to similar problems.

Chapter 4

Transformations

Data for the estimation of heterosis often show heterogeneity of variance, as we shall later see in an example with phenotypic data. Dominance may be estimated by microarray data, which are also known to be extremely heterogeneous concerning variance. Therefore it is frequently necessary to transform either the data or, within the context of generalized linear models, the linear predictor, to satisfy certain assumptions. For microarrays the log-transformation is probably the most common transformation. Other transformations are possible like the so-called generalized logarithm, which was independently introduced for microarrays by Durbin, Hardin, Hawkins, and Rocke (2002), Huber, Von Heydebreck, Sültmann, Poustka, and Vingron (2002) and Munson (2001). This transformation converges to the natural logarithm for high intensities and stabilizes the variance to the first order, meaning that the first order Taylor expansion has constant variance.

In this chapter it will be argued that the amount of heterosis is scale-dependent varying with the kind of transformation. The same applies for the examination of dominance in quantitative genetics. The varying heterotic effect is exemplified using the Box-Cox transformation with phenotypic data of maize roots. Either a data transformation or a generalized linear mixed model with appropriately chosen link function is applied to the data. It is concluded that care should be exercised when transforming data in phenotypic as well as quantitative-genetic studies because partial dominance or heterosis may be removed by a suitably chosen transformation. With data transformations, even overdominance or better parent heterosis may disappear. When a data transformation is needed to meet the usual statistical assumptions such as normality and homogeneity of variance, a back-transformation to the original scale may be necessary, depending on what is deemed the appropriate scale for assessing genetic effects. The findings described in this chapter are also depicted in Keller and Piepho (2005).

4.1 Theory of transformations

Quite often some or all of the assumptions underlying a linear model are not satisfied. While non-normality does not seem to be a major problem with large samples as a result of the central limit theorem, independence and homogeneity of variance are far more important. Transforming the data may be a solution. We consider the Box-Cox transformation as given by Box and Cox (1964):

$$t(y_i;\phi)=\begin{cases}\frac{y_i^{\phi}-1}{\phi} & \text{if } \phi\neq 0\\ \ln(y_i) & \text{if } \phi=0\end{cases}, \tag{4.1}$$

where $t(y_i;\phi)$ is the transformed value and ϕ is a transformation parameter. In order to better meet the usual assumptions, the transformation parameter may be estimated by the Maximum Likelihood (ML) method, assuming normality and homogeneity of variance on the transformed scale (Atkinson, 1985, p. 85). Recently, Gurka, Edwards, and Nylander-French

(2007) developed inference tools for testing the transformation parameter in mixed models against a hypothesized value. This is useful in applied settings, where one is not interested in the exact transformation parameter, but wants to apply a certain preferred value, as, e.g., applying $\phi = 0$ in the case of the Box-Cox transformation. The Box-Cox transformation often gives good results regarding normality. Gurka, Edwards, Muller, and Kupper (2006) showed that when an extended version of the Box-Cox transformation of the response in a mixed model results in near normality of the total error term, the random effects and the residual error will each have approximate normal distributions.

Transformations in general, however, have the disadvantage that a transformation providing normality will not always protect from variance heterogeneity. Therefore, a generalized linear model (GLM) is often preferred. Within this context a wide variety of data may be modeled. The data y_i consists of measurements from a distribution of the exponential family, which is characterized by

$$f_{Y_i}(y_i) = \exp\{\frac{a(y_i)\gamma_i - b(\gamma_i)}{\tau^2} - c(y_i, \tau)\}, \tag{4.2}$$

for some specific functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. The Y_i are assumed to be independent and have expectation μ_i: $E[y_i] = \mu_i$, which is connected to the linear part of the model by a link function $g(\cdot)$. The linear part is denoted linear predictor η_i:

$$\eta_i = x_i'\beta, \tag{4.3}$$

where x_i is the i-th row vector in the design matrix X. Therefore the linear predictor and the expectation of the data are connected by

$$\mu_i = g^{-1}(\eta_i). \tag{4.4}$$

Let's go back to (4.2). Suppose we parametrize the distribution function in a way that $a(y_i) = y_i$, then the parametrization is called canonical and γ_i is sometimes called the natural parameter. According to McCulloch and Searle (2001, p. 140)

$$\mu_i = \frac{\partial b(\gamma_i)}{\partial \gamma_i} \tag{4.5}$$

and

$$\text{Var}(y_i) = \tau^2 \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2} = \tau^2 v(\mu_i),$$

where $v(\mu_i) = \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}$. With (4.5) and (4.4) we see that $g^{-1}(\eta_i) = \frac{\partial b(\gamma_i)}{\partial \gamma_i}$, that is, the derivative of $b(\gamma_i)$ can be the inverse link function for the model. This natural link function is called the canonical link of the distribution. To simplify matters we will use only canonical parametrizations hereafter, but not necessarily canonical link functions. The exponential family includes both continuous and discrete distribution functions as, for example, the Gaussian, Gamma, Poisson and Binomial distribution. These distributions all have a canonical link function. In case of Gaussian data the canonical link function is the identity link $\mu = \eta$.

We consider random variables $Y_i, i = 1, ...n$ that follow the assumptions of a generalized linear model with canonical parametrization, i.e. the Y_i have the following distribution:

$$f_{Y_i}(y_i) = \exp\{\frac{y_i \gamma_i - b(\gamma_i)}{\tau^2} - c(y_i, \tau)\}.$$

In case of a generalized linear mixed model (GLMM) with random effects u, it is not the y_i, but the $y_i|u_i$, i.e. the observations conditional on the random effects, which are distributed with a density from the exponential family:

$$f_{Y_i|u}(y_i|u_i) = \exp\{\frac{y_i \gamma_i - b(\gamma_i)}{\tau^2} - c(y_i, \tau)\}.$$

The random effects are assumed to have a certain distribution $u \sim f_U(u)$. The expectation of y_i is now conditional on the random effects: $E[y_i|u] = \mu_i$ and the connection between μ_i and the linear predictor is $g(\mu_i) = x_i'\beta + z_i'u$, where z_i is the i-th row vector of the design matrix Z of the random effects. The likelihood of the Y_i is obtained by integrating over the random effects:

$$\begin{aligned}
L &= f_{Y_1}(y_1) \cdot \ldots \cdot f_{Y_n}(y_n) \\
&= \int \prod_i f_{Y_i,u}(y_i, u) du \\
&= \int \prod_i f_{Y_i|u}(y_i|u) f_U(u) du \\
&= \int f_{Y|u}(y|u) f_U(u) du.
\end{aligned}$$

Estimation can be performed by maximizing the log-likelihood l. For the random effects, this is given by

$$\begin{aligned}
\frac{\partial l}{\partial \varphi} &= \frac{\partial}{\partial \varphi} \log \int f_{Y|u}(y|u) f_U(u) du \\
&= \frac{1}{\int f_{Y|u}(y|u) f_U(u) du} \frac{\partial}{\partial \varphi} \int f_{Y|u}(y|u) f_U(u) du \\
&= \frac{1}{\int f_{Y,U}(y,u) du} \int \frac{\partial}{\partial \varphi} f_{Y|u}(y|u) f_U(u) du \\
&= \frac{1}{f_Y(y)} \int \left(\frac{1}{f_{Y|u}(y|u) f_U(u)} \frac{\partial f_{Y|u}(y|u) f_U(u)}{\partial \varphi} \right) f_{Y|u}(y|u) f_U(u) du \\
&= \frac{1}{f_Y(y)} \int \frac{\partial \log \left[f_{Y|u}(y|u) f_U(u) \right]}{\partial \varphi} f_{Y|u}(y|u) f_U(u) du \\
&= \int \frac{\partial \log \left[f_{Y|u}(y|u) f_U(u) \right]}{\partial \varphi} \frac{f_{Y,u}(y,u)}{f_Y(y)} du \\
&= \int \frac{\partial \log f_{Y,U}(y,u)}{\partial \varphi} f_{u|y}(u|y) du \\
&= \mathrm{E} \left[\frac{\partial \log f_U(u)}{\partial \varphi} |y \right].
\end{aligned}$$

This equation can be simplified when the distribution of the random effects is known. Similarly to the computations above, the log-likelihood

Table 4.1: *Genotypes and their genotypic values.*

Genotype	Expected genotypic value (μ)	Dummy for additivity effect $\alpha(x)$	Dummy for heterotic effect $\delta(z)$	i
aa	γ	0	0	1
Aa	$\gamma + \alpha + \delta$	1	1	2
AA	$\gamma + 2\alpha$	2	0	3

may be differentiated with respect to the fixed effects β resulting in

$$\frac{\partial l}{\partial \beta} = \int \frac{\partial \log f_{Y|u}(y|u)}{\partial \beta} f_{U|y}(u|y) du.$$

Equating both derivatives to zero leads to the ML-equations. However, in most cases they cannot be solved analytically and numerical quadrature methods are used. In this thesis these computations are performed by adaptive Gaussian quadrature as described by Pinheiro and Bates (1995).

4.2 A model for heterosis and dominance

At the phenotypic level, mid-parent heterosis (MPH) is defined in (1.1) as the superiority of a hybrid compared to the parental mean, whereas better-parent heterosis (BPH, (1.2)) indicates the superiority of a hybrid compared to the better parent. Let us introduce a model for the expected phenotypic value of a certain genotype. Consider the genotypes given in Table 4.1, which may stem, e.g., from a cross Aa of two maize inbred lines with genotypes aa and AA. The linear model for phenotypic values can be stated as

$$\mu_i = \gamma + \alpha x_i + \delta z_i, \tag{4.6}$$

where μ_i is the expected phenotypic value of i-th genotype and $\gamma, \alpha, x_i, \delta$ and z_i are, respectively, the expected phenotypic characteristic of genotype

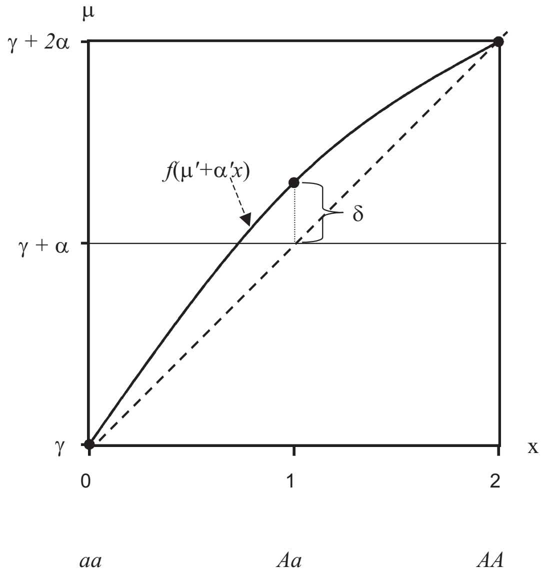


Figure 4.1: *Plot of genotypic value (μ) versus dose (x) of allele A.*

aa, an additive effect (i.e. half the difference of the two parent means), the dose of A for the i-th genotype, the mid-parent heterosis, and the dummy variable for the heterozygote. In this case, α and δ cannot usually be ascribed to the action of a single gene, except when the parents are near-isogenic lines differing in but one locus. Model 4.6 can be visualized by plotting the phenotypic or genotypic value against the dose of A (Figure 4.1). When no mid-parent heterosis is present, i.e., when $\delta = 0$, the model simplifies to

$$\mu_i = \gamma + \alpha x_i.$$

In the context of quantitative genetics a and A may be considered as alleles from a diallelic locus with segregating genotypes AA, Aa and aa in the F$_1$-population. Assuming that a closely linked marker is available, segregation at the locus can be studied directly by comparing marker class means for phenotypic data (Boiteux et al., 2004). Alternatively, the gene expression may be studied by using cDNA microarrays, as described in the previous chapter. In this case, the response variable is a measure of the quantity of expression products in the plant tissue considered. When analysing marker data, the terms $\mu_i, \gamma, \alpha, x_i, \delta$ and z_i in model (4.6) stand for the expected value of the gene expression of the i-th genotype, the expected value of the gene expression of genotype aa, the additive effect of allele A, the dose of allele A for the i-th genotype, the dominance effect, and the dummy variable for the heterozygote. We can distinguish different degrees of dominance such as overdominance, partial dominance and complete dominance. Overdominance is present when $|\delta| > |\alpha|$. When $|\delta| < |\alpha|$, there is only partial dominance, while complete dominance occurs when $|\delta| = |\alpha|$ (Falconer & Mackay, 1996, p. 26). Equivalently, the degree of dominance may be characterized by the dominance ratio $\rho = \frac{\delta}{|\alpha|}$ (Table 4.2). Simultaneous confidence intervals for ρ, α and δ may be calculated according to Piepho and Emrich (2005).

The idea of different degrees of dominance can be carried forward to the heterosis context. Considering the whole genome, one can build the sum of effects over all loci. Some of them might cancel out while others sum up (Mather & Jinks, 1977), and the resulting degree of heterosis can be classified in different groups. If we confine attention to the agronomically interesting cases where MPH > 0, three types of heterosis can be distinguished, according to whether the hybrid performance is less than, equal to or is greater than the performance of the better parent, i.e. BPH > 0, BPH $= 0$ and BPH < 0, respectively. Thus, the modelling of dominance and heterosis data are perfectly congruent (Table 4.2). To sustain the analogy between dominance and heterosis, an additive effect has been included in model (4.6), although it is not commonly used explicitly in

Table 4.2: *Characterization (Ch.) of different degrees of dominance and corresponding degrees of heterosis*

Degree of dominance (d.)	Corresponding degree of heterosis[§]	Ch. in terms of δ and α	Ch. in terms of $\rho = \delta/\lvert\alpha\rvert$
Overd.	BPH>0, MPH>0	$\delta > \lvert\alpha\rvert$	$\rho > 1$
Complete d.	BPH$=0$, MPH>0	$\delta = \lvert\alpha\rvert$	$\rho = 1$
Partial d.	BPH<0, MPH>0	$\delta < \lvert\alpha\rvert$	$0 < \rho < 1$
Lack of d.	BPH<0, MPH$=0$	$\delta = 0$	$\rho = 0$

[§] MPH = mid-parent heterosis; BPH = better-parent heterosis. Without loss of generality we assume that MPH$\geq$0, so that $\delta \geq 0$.

phenotypic analysis of heterosis. The effect δ, however, is common in both contexts and can be interpreted either as mid-parent heterosis, when looking at a phenotypic trait, or dominance, when looking at one locus. The relationship between heterosis at a single locus and dominance is also shown in Bernardo (2002, p. 243).

4.3 Influence of transformations on heterosis estimates

The analysis of the linear model (4.6) by standard procedures may be based on the usual assumptions such as additivity, homogeneity of variance, and normality. When at least one of these assumptions is violated, one may avail oneself of the methods proposed in Section 4.1. The most common reaction is to search for a data transformation, which will meet all assumptions simultaneously, at least approximately. This approach has been used frequently in studies of heterosis and dominance (Boiteux et al., 2004; Baker et al., 2003; Tefera & Peat, 1997; Roumen, 1994). Alternatively, one may take recourse to a generalized linear model analysis (McCullagh & Nelder, 1989), which transforms the linear predictor rather than the data. In the GLM context, this transformation is also known as the link function. In either case, the transformation will not leave the heterotic or dominance effect unaffected. Specifically, one can always find a data transformation or a link function that makes this effect disappear, providing the BPH on the original scale is smaller than zero. In case of a data transformation, even positive BPH may disappear. These important facts will be demonstrated by examples using phenotypic data. However, due to the above-mentioned analogy between heterosis and dominance, the same problem exists for dominance on the original scale. The implications are twofold. Firstly, in quantitative-genetic studies care should be exercised when transforming data or linking a linear predictor. It should be critically checked, whether the transformed scale is useful or natural for studying heterotic (dominance) effects. Secondly, if a transformation is needed only to better meet the statistical assumptions, one should back-transform parameter estimates to the original scale for inference on genetic effects. It will be shown that the generalized linear mixed model framework offers flexibility to account for non-normality and variance heterogeneity, so that analysis can focus on a transformation (link function) that is deemed optimal for the study of genetic effects.

4.3.1 Theoretical approach

In this section it will be shown that partial dominance or heterosis may be removed by a generalized linear model with a suitably chosen transformation parameter. In analogy to (4.1) we use the link function

$$\eta_i = g(\mu_i; \phi) = \begin{cases} \frac{\mu_i^\phi - 1}{\phi} & \text{if } \phi \neq 0 \\ \ln(\mu_i) & \text{if } \phi = 0 \end{cases} .$$

With the indices from Table 4.1 mid-parent heterosis on the transformed scale can be described as $\delta(\phi) = \eta_2 - \frac{1}{2}(\eta_1 + \eta_3)$. In case of negative BPH of the hybrid it can be assumed without loss of generality that on the original scale $0 < \mu_1 < \mu_2 < \mu_3$ and means can be re-expressed as

$$\begin{aligned} \mu_1 &= \theta \\ \mu_2 &= \theta\lambda_1 \\ \mu_3 &= \theta\lambda_1\lambda_2 \end{aligned}$$

with $\theta > 0, \lambda_1 > 1$, and $\lambda_2 > 1$.

Heterosis on the transformed scale is given by

$$\delta(\phi) = \begin{cases} \frac{\theta^\phi}{\phi}\left(\lambda_1^\phi - \frac{1+\lambda_1^\phi\lambda_2^\phi}{2}\right) & \text{if } \phi \neq 0 \\ \frac{1}{2}[\ln(\lambda_1) - \ln(\lambda_2)] & \text{if } \phi = 0 \end{cases} . \quad (4.7)$$

In the following we confine attention to cases where MPH is present for untransformed data, i.e., where $\theta\lambda_1 - \frac{1}{2}(\theta + \theta\lambda_1\lambda_2) > 0$. This is equivalent to

$$2 - \frac{1}{\lambda_1} > \lambda_2. \quad (4.8)$$

As $-(\lambda_1 - 1)^2 \leq 0$ holds true for all λ_1, simple calculus leads to $2 - \frac{1}{\lambda_1} \leq \lambda_1$. Together with (4.8) we come to the result that for positive MPH $\lambda_1 > \lambda_2$.

MPH disappears when $\delta(\phi) = 0$. For $\phi = 0$ this is not possible, as we confined ourselves to cases where $\lambda_1 > \lambda_2$. If $\phi \neq 0$, for MPH to disappear the following condition must be fulfilled:

$$\lambda_1^\phi - \frac{1 + \lambda_1^\phi\lambda_2^\phi}{2} = 0 \qquad \Longleftrightarrow \qquad 2 - \lambda_2^\phi = \lambda_1^{-\phi}.$$

Solutions for this equation other than the trivial case $\phi = 0$ (which is not a solution to $\delta(\phi) = 0$) can only be determined numerically, e.g., by Newton's method. We now show that there is always exactly one such solution.

Let $g_1(\phi) = \lambda_1^{-\phi}$ and $g_2(\phi) = 2 - \lambda_2^{\phi}$. The first two derivatives with respect to ϕ are found to be

$$\begin{aligned} g_1'(\phi) &= -\lambda_1^{-\phi}\ln(\lambda_1) \\ g_1''(\phi) &= \lambda_1^{-\phi}[\ln(\lambda_1)]^2 \\ g_2'(\phi) &= -\lambda_2^{\phi}\ln(\lambda_2) \\ g_2''(\phi) &= -\lambda_2^{\phi}[\ln(\lambda_2)]^2 \end{aligned}$$

The first derivative of both functions is negative for all ϕ, so the functions are monotonically decreasing in ϕ. The second derivative of both functions has the same sign for all ϕ. A function $f(\phi)$ is said to be convex (concave) if its second derivative is positive (negative) for any value of ϕ. It is found that g_1 is convex and g_2 is concave. This is sketched in Figure 4.2. Moreover,

$$\begin{aligned} &\lim_{\phi\to\,-\infty} g_1(\phi) = \infty \quad \text{and} \quad \lim_{\phi\to\infty} g_1(\phi) = 0 \\ &\lim_{\phi\to\,-\infty} g_2(\phi) = 2 \quad \text{and} \quad \lim_{\phi\to\infty} g_2(\phi) = -\infty. \end{aligned}$$

Considering these facts it can be said that the two functions may have one of three possible joint patterns: (i) the curves do not intersect; (ii) the curves touch in one point; (iii) the curves intersect at two points. We already found that the functions always meet at $\phi = 0$ (a trivial case), so pattern (i) can be ruled out. Gradients of $g_1(\phi)$ and $g_2(\phi)$ at $\phi = 0$ are different as $\lambda_1 > \lambda_2$, so that pattern (iii) must apply, while pattern (ii) can be ruled out as well. Hence, there must always be a second (non-trivial) point of intersection, which we denote by ϕ_0. From $\lambda_1 > \lambda_2$ follows that $-\ln(\lambda_1) < -\ln(\lambda_2)$, i.e., the gradient of g_1 in $\phi = 0$, the convex curve, is smaller than the gradient of g_2, the concave curve, which means that the non-trivial point of intersection between the two is for $\phi_0 > 0$.

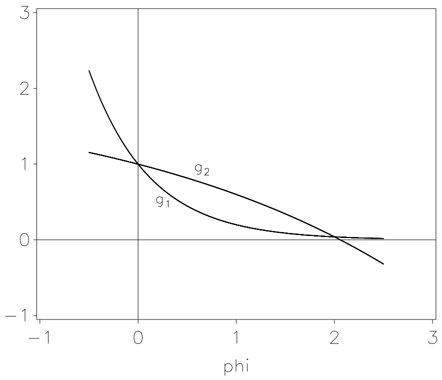


Figure 4.2: $g_1(\phi)$ *and* $g_2(\phi)$ *for* $\lambda_1 = 5$ *and* $\lambda_2 = 1.4$

To sum up, for a hybrid with positive MPH in the raw data there is exactly one parameter value ϕ that removes the heterotic effect. It is, however, not only possible to reduce or even eliminate the heterotic effect, it can also be enlarged.

Generally, we are not only interested in the absolute value of δ, but also in the dominance ratio $\rho = \frac{\delta}{|\alpha|}$, which sets the dominance effect in relation to the additivity effect. With the parametrization above, additivity on the transformed scale is given by

$$\alpha(\phi) = \begin{cases} \frac{\theta^\phi}{2\phi}(\lambda_1^\phi \lambda_2^\phi - 1) & \text{if } \phi \neq 0 \\ \frac{1}{2}[\ln(\lambda_1) + \ln(\lambda_2)] & \text{if } \phi = 0 \end{cases} .$$

As $\alpha(\phi) > 0$ for all ϕ, we have $\alpha_\phi = |\alpha_\phi|$ and with (4.7) the dominance ratio on the transformed scale is

$$\rho(\phi) = \frac{\delta(\phi)}{\alpha(\phi)} = \begin{cases} \frac{2\lambda_1^\phi - 1 - \lambda_1^\phi \lambda_2^\phi}{\lambda_1^\phi \lambda_2^\phi - 1} & \text{if } \phi \neq 0 \\ \frac{\ln(\lambda_1) - \ln(\lambda_2)}{\ln(\lambda_1) + \ln(\lambda_2)} & \text{if } \phi = 0 \end{cases} . \tag{4.9}$$

Using l'Hospital's rule we have $\lim_{\phi \to 0} \rho(\phi) = \frac{\ln(\lambda_1) - \ln(\lambda_2)}{\ln(\lambda_1) + \ln(\lambda_2)}$, and thus (4.9) is a continuous function. We determine the limits of ρ_ϕ for $\phi \to \pm\infty$:

$$\lim_{\phi \to -\infty} \rho(\phi) = \lim_{\phi \to -\infty} \frac{2\lambda_1^\phi - 1 - (\lambda_1 \lambda_2)^\phi}{(\lambda_1 \lambda_2)^\phi - 1} = 1$$

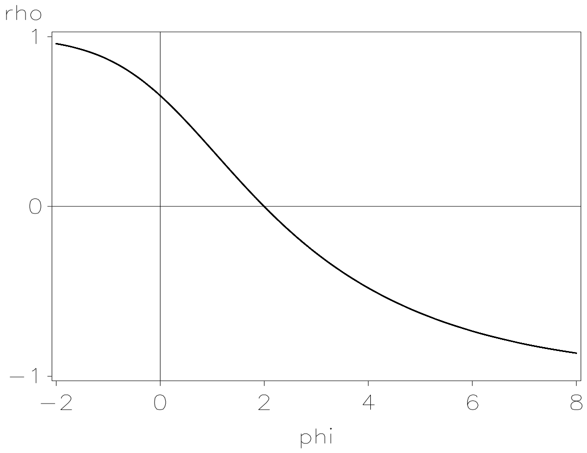


Figure 4.3: $\rho(\phi)$ *for* $\lambda_1 = 5$ *and* $\lambda_2 = 1.4$

and

$$\lim_{\phi\to\infty} \rho(\phi) = \lim_{\phi\to\infty} \frac{2\lambda_1^{\phi} - 1 - (\lambda_1\lambda_2)^{\phi}}{(\lambda_1\lambda_2)^{\phi} - 1} = \lim_{\phi\to\infty} \frac{2\lambda_2^{-\phi} - (\lambda_1\lambda_2)^{-\phi} - 1}{1 - (\lambda_1\lambda_2)^{-\phi}} = -1.$$

In Figure 4.3 $\rho(\phi)$ is plotted against ϕ for $\lambda_1 = 5$ and $\lambda_2 = 1.4$. It can be shown that $\rho(\phi)$ is monotonically decreasing (Appendix A). Together with the consideration of the limits, this has the following interpretation: for untransformed data $\rho(\phi = 1)$ must be positive as we assume that MPH is present and that the parents are different. Performing a Box-Cox transformation with $\phi > 1$ the dominance ratio is reduced. It may become zero or even negative with lower bound minus one. Transforming with a parameter value $\phi < 1$, the dominance ratio is enlarged, but it will not reach one.

We showed that mid-parent heterosis and the dominance ratio can always be turned to zero by a suitably chosen transformation of the linear predictor. For the dominance ratio we showed that it may also be enlarged by a transformation. Contrary to transformations when MPH is present, positive BPH cannot be removed by a monotone transformation of the linear predictor, and the distinction among positive BPH and other forms of heterosis (Table 4.2) will not be affected by a monotone transformation of the linear predictor. To see this, consider the contrasts $\delta - \alpha$ and $\delta + \alpha$. It is easily verified that $|\delta| > |\alpha|$ only if both of these contrasts have equal sign. Observing that $\delta + \alpha = \eta_2 - \eta_1$ and $\delta - \alpha = \eta_2 - \eta_3$, it is clear that a monotone transformation of η_i will have no effect on the decision whether or not $|\delta| > |\alpha|$ holds true, since ranking of η_i is not altered by a monotone transformation.

4.3.2 Practical approach using experimental data

The effects of transformations are demonstrated by means of phenotypic data from early maize seedlings as described in Section 2.1.4. To illustrate the effects of transforming either the original data or the linear predictor, the hybrid UH005×UH301 was chosen. The hybrids performance lies between the parental mean and the better parent (i.e. MPH> 0, BPH< 0) on the original scale. Hybrid UH250×UH005 was selected to illustrate that even positive BPH may be removed by data transformations in some cases.

The data are analysed based on an extension of model (4.6), which has the form

$$t(y_{ijk}) = \gamma + \alpha x_i + \delta z_i + p_{ij} + e_{ijk},$$

where $t(\cdot)$ is a transformation, y_{ijk} the length of the k-th lateral root of the i-th genotype class and j-th primary root, p_{ij} is the random effect of the j-th primary root in genotype class i and e_{ijk} is an error term. Both p_{ij} and e_{ijk} are assumed to follow a normal distribution with zero mean and homogeneous variance.

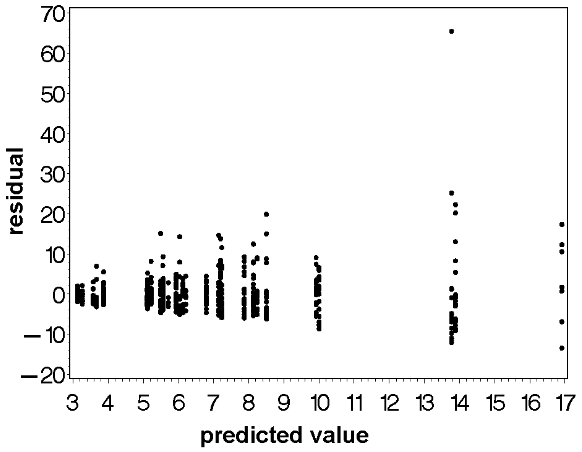


Figure 4.4: *Plot of residuals vs. the predicted values show that neither normality nor variance homogeneity are satisfied.*

Plots of the predicted values vs. the residuals of the untransformed data y_{ijk} show an increase of variance with the prediction (Figure 4.4). Therefore, the assumption of homogeneous variance is violated for the original data y_{ijk}, and transformation of either the data or the linear predictor is necessary.

Transforming the data

The data are transformed with the Box-Cox transformation (4.1) where the transformation parameter ϕ is estimated by the Maximum Likelihood method. However, if a value for ϕ is chosen that differs from the ML estimate, estimates for δ and other effects are different. For example, for a hybrid with negative BPH on the original scale, one can find a parameter value for ϕ that removes MPH on the transformed scale. Therefore, the ML estimate was determined as well as the parameter value that minimizes the F-statistic for the effect for δ and the resulting estimates for α, δ and ρ were calculated. In addition, estimates for the often-used log-transformation were determined, as well as estimates of untransformed data, although these are not reliable as the assumptions of normality and variance homogeneity are not satisfied.

If the objective of the data transformation is merely to meet certain assumptions, results should be backtransformed to the original scale. The delta method, which is based on a first-order Taylor-series expansion, may be used to compute an approximation for the expectation on the original scale (Lynch & Walsh, 1998). However, this is only a good approximation if the variance in the original data is low. We therefore prefer to determine the median on the original scale. As we act on the assumption of normality on the transformed scale, the mean on this scale is an estimate of the median. Due to the monotony of the transformations the data are ordered the same way on the original and on the transformed scale. Hence it is straightforward to compute the median on the original scale by applying the inverse transformation to the mean of the transformed data (Connolly & Wachendorf, 2001). Apart from computational simplicity, the median is preferable to the mean as a location measure in case of skewedly distributed data. Estimates for logarithmized data and ML-estimates are backtransformed to medians on the original scale, while for the estimates minimizing the F-statistic for dominance (see below) no backtransformation is done as this transformation was only conducted to show the disappearance of δ and will not be used in practice.

Transforming the linear predictor

In the preceding section it has been pointed out that heterosis can often be removed by a data transformation. In the present section, we will show that, within the framework of generalized linear mixed models (Section

4.1), one can always find a monotone transformation $g(\mu)$ of the expectation μ of the vector of observations that removes heterosis when BPH< 0 and MPH> 0.

For the estimation of heterotic effects the linear predictor η from (4.3) may be expressed as

$$\eta_i = g(\mu_i) = \gamma + \alpha x_i + \delta z_i,$$

with $\gamma, \alpha, \delta, x_i$ and z_i defined as in model (4.6). We let the random effect of the primary roots and the residual error enter the model as follows:

$$y_{ijk} = g^{-1}(\eta_i) + p_{ij} + e_{ijk}. \tag{4.10}$$

This is known as a population-averaged model, as opposed to a subject-specific model, where p_{ij} would enter the model via the linear predictor (Vonesh & Chinchilli, 1997; Schabenberger & Pierce, 2002, p. 416). The claim here is that with heterosis such that MPH> 0 and BPH< 0 on some original scale, one can always find a monotone transformation $g(\mu_i)$ involving a transformation parameter ϕ so that heterosis vanishes, i.e.,

$$g'(\mu_i) = \gamma' + \alpha' x_i$$

on the transformed scale, where α' is the additive genetic effect for A and γ' the genotypic value for genotype aa. Thus, one can always find a link function that removes heterosis when MPH> 0 and BPH< 0 (4.3.1) and the presence and magnitude of negative BPH is entirely scale-dependent. This is illustrated calculating estimates by means of the GLMM framework with different link functions: identity link, log-link and Box-Cox link. The parameter of the latter is chosen in a way that minimizes the Wald-type F-statistic for the effect δ. The Box-Cox link is just an example of a link function removing heterosis. This link function involves the parameter

ϕ, which requires special attention when fitting a GLMM (McCullagh & Nelder, 1989, p. 375). For all three models, i.e. identity, log- and Box-Cox link, the error e_{ijk} is assumed to be gamma distributed, while p_{ij} follows a normal distribution. The gamma distribution includes a scale parameter allowing the modelling of a variance that increases with the expectation, which is in agreement with the variance pattern we observed for the root data (Figure 4.4).

4.4 Results and discussion

In Section 4.3.1 it was shown that partial heterosis or dominance may be enlarged or reduced by a suitably chosen transformation of the linear predictor in a generalized linear model. In particular partial heterosis or dominance may be completely removed, i.e., the effect for heterosis or dominance becomes zero. For our practical approach the hybrid UH005×UH301 and the parents were considered. Root length of the hybrid showed negative BPH on the original scale. Estimates of genotype effects, additive effect (α), heterotic effect (δ) and heterosis-additivity ratio ρ for different data transformations as well as for untransformed data (Table 4.3) were determined.

With untransformed data, the MPH δ is very low. However, these estimates are not reliable as the untransformed data have heterogeneous variance and residuals are not normally distributed. When we assume the response variable to be lognormally distributed, the heterosis-additivity ratio ρ is higher. Performing a Box-Cox transformation the log-likelihood has a clear maximum (Figure 4.5a) at the ML estimate.

Table 4.3: *Estimates for maize hybrid UH005×UH301 and parental inbred lines on original scale and on transformed scales for different data transformations*

	Data analysed							
	Untransformed[§]		log		Box-Cox$_{\mathrm{ML}}^{\dagger}$		Box-Cox$_{\mathrm{Fmin}}^{\ddagger}$	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Genotype mean								
P1 (UH301)	2.78	0.66	0.78	0.11	0.73	0.10	1.85	0.72
F1 (UH005×UH301)	4.94	0.68	1.35	0.11	1.25	0.11	4.15	0.75
P2 (UH005)	7.01	0.81	1.55	0.13	1.41	0.13	6.44	0.89
Genetic effects								
α	2.11	0.52	0.39	0.08	0.34	0.08	2.29	0.57
δ	0.04	0.86	0.19	0.14	0.18	0.12	0.00	.
ρ	0.02	0.41	0.48	0.39	0.52	0.39	0.00	.

[§] Lateral root length [mm].

[†] Estimated by Maximum Likelihood, $\hat{\phi}_{\mathrm{ML}} = -0.10$.

[‡] Estimated by minimizing the F-statistic for δ, $\hat{\phi}_{\mathrm{Fmin}} = 1.04$.

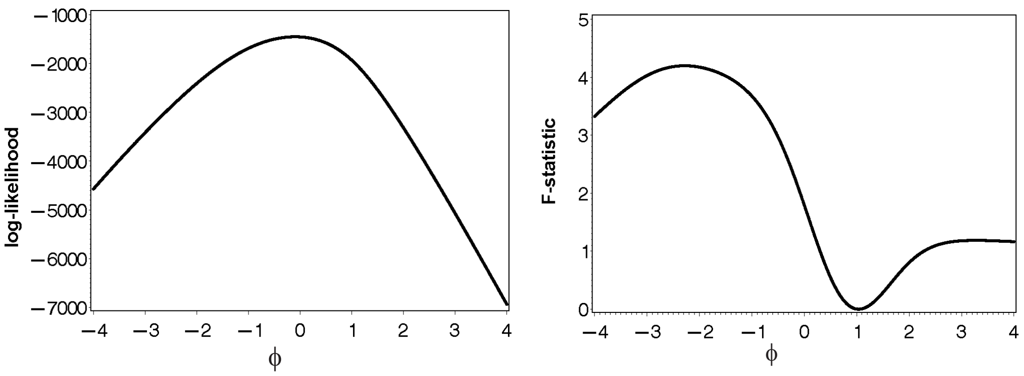


Figure 4.5: *(a) Profile likelihood of f for the data transformation to normality and (b) profile of f for the F-statistic for the dominance effect.*

Transformation with this parameter value gives again a different result. Furthermore, with a Box-Cox transformation parameter close to unity, the F-statistic for the heterotic effect is minimum and heterosis disappears almost completely. This can be visualized by plotting the value of the F-statistic against the transformation parameter ϕ (Figure 4.5b). Of course, one would not choose this parameter value in practice, as a Box-Cox parameter of unity corresponds to untransformed data, for which the model assumptions are not met. However, it can be seen clearly (Figure 4.6) that the heterosis-additivity ratio ρ is influenced by the choice of value for the transformation parameter.

Backtransformation shows that with the log-transformation as well as the Box-Cox-transformation estimates of α and δ are smaller, whereas the estimates of the heterosis-additivity ratio are higher (Table 4.4).

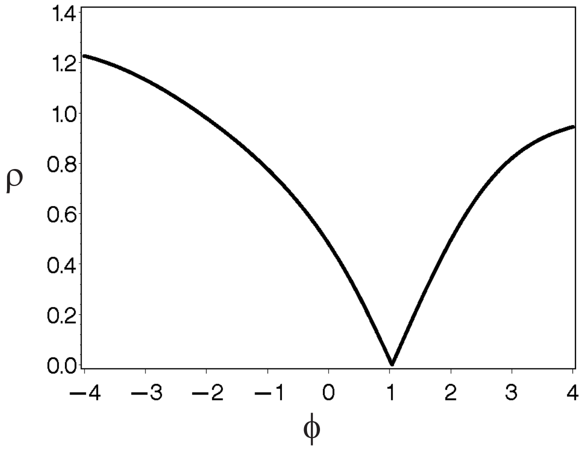


Figure 4.6: *Different values of Box-Cox transformation parameter ϕ result in different estimates of dominance ratio ρ.*

Table 4.4: *Backtransformed estimates for maize hybrid UH005×UH301 and parental inbred lines for log- and Box-Cox transformed data*

	Data analysed			
	log		Box-Cox$_{\mathrm{ML}}$	
	Estimate	S.E.	Estimate	S.E.
Genotype mean				
P1 (UH301)	2.17	0.23	2.12	0.22
F1 (UH005×UH301)	3.86	0.43	4.56	0.62
P2 (UH005)	4.71	0.62	3.77	0.42
Genetic effects				
α	1.27	0.33	1.22	0.33
δ	0.41	0.54	0.43	0.53
ρ	0.33	0.47	0.35	0.49

The fact that heterotic effects can be removed by a data transformation may also show up with hybrids where the hybrid's performance exceeds the better parent (positive BPH). The reason for this is that means are computed across transformed data. As an example, we analysed transformations for hybrid UH250×UH005. When estimating effects for genotypes, means are computed over the transformed data. Thus it was possible to find a Box-Cox parameter so that the value of the F-statistic for δ becomes negligible. In Figure 4.7 this result is shown for genotype effects. The least-squares mean of the hybrid lies noticeably higher than the least squares means of both parents when analyzing raw data. If the Box-Cox transformation with transformation parameter value ϕ_{Fmin} is performed before the analysis, the least-squares mean of the hybrid lies in between the least-squares means of the parents. Estimates and standard errors of the three genotypes are indicated in Table 4.5. It should be noticed that standard errors for the Box-Cox transformed data are extremely high.

The second approach was to fit generalized linear models with different link functions to the data. Depending on the link function the estimates are quite different (Table 4.6). Again one can find a parameter of the Box-Cox transformation of the linear predictor that removes the heterotic effect. This agrees with our findings from the theoretical approach in Section 4.3.1.

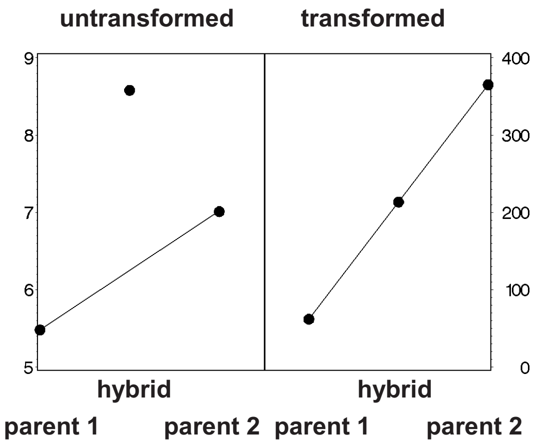


Figure 4.7: *Even overdominance may be removed by a data transformation (least-squares means for parent 1: UH250; parent 2: UH005; hybrid: UH250×UH005). Data were transformed using the Box-Cox-transformation with* $\hat{\phi}_{F\,\min} = 2.58$ *(see Table 4.3).*

Table 4.5: *Estimates for maize hybrid UH250×UH005 and parental inbred lines on original scale and Box-Cox transformed data*

	Data analysed			
	Untransformed[§]		Box-Cox$^{\dagger}_{F\,\min}$	
	Estimate	S.E.	Estimate	S.E.
Genotype mean				
P1 (UH250)	5.48	1.17	61.74	137.52
F1 (UH250×UH005)	8.58	1.13	213.47	143.40
P2 (UH005)	7.01	1.02	365.17	139.25
Genetic effects				
α	0.77	0.78	151.72	97.86
δ	2.34	1.37	0.01	.
ρ	3.04	3.42	0.00	.

§ Lateral root length [mm].

† Estimated by minimizing the F-statistic for δ, $\hat{\phi}_{F\,\min} = 2.58$.

Table 4.6: *Estimates for maize hybrid UH005×UH301 and parental inbred lines on original scale and on transformed scale based on generalized linear model*

	Link function $\eta_i = g(\mu_i)$					
	μ_i		$\log(\mu_i)$		Box-Cox[§]	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Genotype mean						
P1 (UH301)	2.82	0.46	1.04	0.16	4.07	1.65
F1 (UH005×UH301)	5.00	0.52	1.61	0.10	15.79	3.76
P2 (UH005)	6.37	0.64	1.85	0.10	27.52	6.30
Genetic effects						
α	1.78	0.39	0.41	0.10	11.72	3.25
δ	0.40	0.65	0.16	0.14	0.00	.
ρ	0.23	0.38	0.40	0.33	0.00	.

[§] Estimated by minimizing the F-statistic for δ, $\hat{\phi}_{F\,\min} = 2.23$.

This work has shown that heterosis is dependent on the choice of scale. Heterotic effects with MPH> 0 and BPH< 0 may be removed by a data transformation as well as a transformation of the linear predictor in the GLM context. Besides the reduction of the heterotic effects, they may also be enlarged by a transformation. If positive better-parent heterosis is present, in many cases a data transformation can also remove this effect, however, with a transformation of the linear predictor this is not possible. To show this theoretically is relatively easy for the transformation of the linear predictor (as seen in Section 4.3.1). In case of data transformations giving a theoretical condition when heterotic effects can be removed is more challenging and has not been examined in this study. As a result of the analogy between estimation of heterosis and dominance similar conclusions can be made for dominance: partial dominance may be removed and enlarged by both a data transformation and a transformation of the linear predictor. Overdominance may sometimes be removed by a data transformation. When a population-averaged GLMM or a GLM with a single error term is used, this will not be possible. A subject-specific GLMM to some extent behaves like a data transformation, because random effects enter the linear predictor, so occasionally it may be possible to remove overdominance.

The disappearance of heterosis is shown by means of phenotypic maize root data. In order to discern the processes underlying the phenomenon of heterosis, one may study the mode of dominance at the gene level. Microarrays may be used to determine the expression level of different genotypes for a great number of genes. Data from expression studies are often logarithmically transformed (Dudoit et al., 2002). This raises the question if the logarithmic scale is a natural scale to study heterosis or dominance at the expression level. We think that the answer to this question is 'yes', when the complex quantitative trait under study is related to simpler component traits in a multiplicative fashion. For example, agronomic yield, perhaps the most important trait for which heterosis is exploited, is the product of yield components, and heterosis in yield has often been found to occur due to multiplicative effects of the component traits (Sparnaaij & Bos, 1993; Melchinger, Singh, Link, Utz, & Kittlitz, 1994; Piepho, 1995; Sant et al., 1999). Conversely, if gene effects on the complex trait are deemed additive rather than multiplicative on the original scale, it may be useful to back-transform mean estimates of expression level using, e.g., the methods proposed in this chapter. The Box-Cox transformation was chosen as an example of transformations influencing heterosis estimates. Applying the generalised logarithm (Durbin et al., 2002) or other transformations the same problem will occur.

The results presented in this paper suggest that generally great care should be exercised when using transformations in phenotypic as well as quantitative-genetic studies. Specifically, due consideration should be given to the question of what constitutes a natural scale on which to assess heterosis or dominance. With count data of the Poisson-type, it seems rather natural to use a log transformation, while with percentage data, it is more natural to use a logit or probit transformation (McCullagh & Nelder, 1989, p. 32). Generally, one may either transform the data or the fitted values, leading to a generalized linear model. A disadvantage of data transformations is that the transformed data need to meet the usual assumptions of normality and homogeneity of variance. It is not generally the case that the natural scale on which to study dominance effects is also the best scale to achieve normality and homogeneity of variance. By contrast, a GLM framework allows one to chose the transformation solely based on the natural scale for studying dominance effects, while distributional assumptions can be relaxed (McCullagh & Nelder, 1989).

Chapter 5

Estimating variance components

A microarray slide contains several thousand spots, which correspond to genes. Usually the data is analysed separately for each gene. As the scientist is seeking information about as many genes as possible, quite often there are no technical replicates on the array, i.e. each gene is spotted only once. Due to financial constraints the number of slides in an experiment is typically low. This means we have a comparatively small number of observations that can be used to estimate genetic effects. It can therefore be difficult to provide evidence that a gene is differentially expressed in different genotypes or tissue types.

What can we do to increase the power of the tests? One possibility is to improve the accuracy of variance components estimates. As a high number of genes is available, the spotwise analysis might not be optimal. Information about the variance components could be lent from other spots, i.e. the variance components could be determined in a joint analysis for all spots (Lönnstedt & Speed, 2002; Smyth, 2004; Gottardo, Raftery, Yeung, & Bumgarner, 2006).

A simple approach is to compute variance components estimates for all spots separately and calculate the mean of the components over spots.

This pooled estimate may be plugged into the gene-specific model. According to Wright and Simon (2003), however, tests based on these variance estimates show a high false positive rate in simulations. This can be ascribed to the fact that highly variable genes may appear as differentially expressed when using a too small variance estimate. Wright and Simon (2003) propose a so-called random variance model, which constitutes a two-stage hierarchical model. Individual variances are assumed to follow an inverse Gamma distribution, while conditionally on the true variances, their estimates have a scaled χ^2-distribution. The estimates are obtained as empirical Bayes estimates according to the estimated hierarchical model. It is shown that tests based on the adjusted variance estimates have higher power than tests based on conventional spotwise variance estimates. However, the approach only works with one variance component. The method used by Cui, Hwang, and Qiu (2005) follows the James-Stein-concept, where individual estimates are shrunken towards a common mean. However, their approach suffers from the same deficiency as the model in Wright and Simon (2003): it cannot be applied to models with more than one variance component. Gottardo et al. (2006) present a fully Bayesian approach to find differential genes between two samples. They assume a multiplicative relation between the array and residual error term. The residual errors of the two samples are supposed to follow a bivariate normal distribution, while the array effect is Gamma distributed. The effect of a sample on a specific gene is modeled as a mixture of two normal distributions, one corresponding to genes that are not differentially expressed, while the other corresponds to differentially expressed genes. Another Bayesian approach proposed by Lewin, Richardson, Marshall, Glazier, and Aitman (2006) assumes the expression values of each sample to be distributed according to an ANOVA model with effects for the overall expression level of the gene, the array effect and an effect for the differential expression between two samples. The expression values are normally distributed with different variances for the two samples. These variances are assumed to follow a lognormal distribution.

In cDNA microarray analysis it is useful to consider the array effect as random when the number of treatments or genotypes is larger than two (Chapter 3). With arrays considered as incomplete blocks, the recovery of inter-block information becomes possible (John & Williams, 1995, p. 27). This may lead to more precise estimates of genetic effects, thus increasing the power of significance tests. Also, when estimating heterotic effects at least three genotypes are involved (the hybrid and both parental lines) and it is often worthwhile to use the inter-array information (Chapter 3). Therefore, we seek for an alternative approach, where at least two variance components, for array and residual variance, are estimated by utilizing all spots on the array. In complex settings, it may be necessary to add further variance components for other random sources of error (Piepho, Büchse, & Emrich, 2003). Here, the focus will be on the case of two variance components, but extension to more than two variance components is straightforward.

5.1 Theory

We introduce a method that provides estimates of array and residual variation that are based on information of all spots on the arrays. We call them 'pooled' variance estimates. For this purpose spotwise estimates of the variance components could be determined and a distribution fit across spots. Common REML-estimates of the array variance may be estimated as zero, however, which makes the fitting of a distribution difficult. We thus employ ANOVA sum of squares of both array and residual effects. This avoids the problem of estimates of zero when fitting a hyper-distribution for the variance components. The distribution of the sum of squares is used to obtain the pooled variance estimates via an empirical Bayes approach, as described below.

A mixed model for each gene is given by

$$y = X\beta + Zu + e, \tag{5.1}$$

where y is the vector of observed signal values on the log scale, β is a vector of fixed effects, u is a vector of random effects, and e is a vector of residual error. We assume here for simplicity that only one random effect exists, besides the residual variation. X and Z are known design matrices for the fixed and random effects, respectively. The components of u and e are independent and follow the normal distribution with mean zero and variances θ_1 and θ_0, respectively. The covariance between elements of u and e is zero. We indicate the variance components for a specific gene j with an index, the true variance components for e, u being $\theta_j = (\theta_{0j}, \theta_{1j})$, for $j = 1, ...J$.

We assume the variance components θ_j to follow a bivariate lognormal distribution with mean vector μ and covariance matrix Σ on the log-scale. This is a strong assumption and other distributions are possible, like the inverse Gamma (Wright & Simon, 2003) or the bivariate Johnson System of transformations (Johnson, 1949) with the lognormal as a special case. However, as a lognormal variable is positive by definition, the lognormal distribution seems to be a natural choice and has been applied successfully in other studies (Cui et al., 2005; Lewin et al., 2006). Also, our former analyses of microarray data within the priority program support the assumption of a lognormal distribution (results not shown). The parameters μ and Σ constitute the hyperparameters of the prior for the variance components. The lognormal prior is denoted by $g(\theta_j|\mu, \Sigma)$.

Let SS_{ij} and MS_{ij} denote the sum of squares and mean squares of variance component i and spot j, using Henderson's method III for the calculation of sum of squares where random effects are fit after fixed effects (Searle et al., 1992, p. 202). This is a classical ANOVA-method, as opposed to the ML- and REML-approaches presented in Chapter 3. Our empirical Bayes approach assumes a χ^2- distribution for the sum of squares of variance components given the true variance components, divided by the expected mean squares:

$$\frac{\mathrm{SS}_{ij}|\theta_j}{\mathrm{E}(\mathrm{MS}_{ij})} \sim \chi^2_{\nu_{ij}}, \tag{5.2}$$

where θ_j are the true variance components and ν_{ij} is the number of degrees of freedom associated with SS_{ij}. It is further assumed that SS_{0j} and SS_{1j} are stochastically independent for given variance components θ_j. It must be stated, however, that in the case of more than two variance components the sum of squares are no longer necessarily independent (Milliken & Johnson, 1992, p. 252), so extension to more than two variance components requires some form of approximation or fitting a multivariate distribution. The expected mean squares are $\mathrm{E}(\mathrm{MS}_{0j}) = \theta_{0j}$ and $\mathrm{E}(\mathrm{MS}_{1j}) = \theta_{0j} + c\theta_{1j}$. The coefficient c is dependent on the spot's design and may be determined by

$$c = \frac{tr(Z'MZ)}{r[X \quad Z] - r[X]},$$

where $M = I - X(X'X)^{-}X'$ (Searle et al., 1992, p. 204). The likelihood of the observed sum of squares given the unknown variance components specified in (5.2) is denoted by $f(\mathrm{SS}_{ij}|\theta_j, \mu, \Sigma)$. Let $\eta = (\mu, \Sigma)$ denote the set of hyperparameters of the prior. The posterior distribution of the true variance components is thus given by

$$\pi(\theta_j|\mathrm{SS}_{ij}, \eta) = \frac{f(\mathrm{SS}_{ij}|\theta_j, \eta)g(\theta_j|\eta)}{\int f(\mathrm{SS}_{ij}|\theta_j, \eta)g(\theta_j|\eta)d\theta_j}. \tag{5.3}$$

Estimates $\hat{\eta}$ of the hyperparameters may be derived using Maximum Likelihood on the marginal distribution $m(\mathrm{SS}_i|\eta)$

$$m(\mathrm{SS}_i|\eta) = \int \prod_{j=1}^{J} f(\mathrm{SS}_{ij}|\theta_j, \eta)g(\theta_j|\eta)d\theta_j. \tag{5.4}$$

For the computation of (5.3) and (5.4), integrals over the random effects have to be evaluated. The Gaussian quadrature approximates an integral by a weighted sum of function values at so-called quadrature points on the abscissa, which are centered around zero. A slightly different approximation method is the adaptive Gaussian quadrature, where quadrature points are not centered around zero, but around the modes of the random effects. It is recommended to use the adaptive Gaussian quadrature as it produces more accurate results and is computationally more efficient (Pinheiro & Bates, 1995; SAS Institute Inc., 1999).

Empirical Bayes (EB) estimates of θ_j are obtained by substituting $\hat{\eta}$ for η in $\mathrm{E}(\theta_j|\mathrm{SS}_{ij}, \eta)$ to give $\hat{\theta}_j = \mathrm{E}(\theta_j|\mathrm{SS}_{ij}, \hat{\eta})$. The estimated variance of the EB estimates is computed as the inverse Hessian matrix. The easiest way to obtain the estimated covariances of array and residual variance using the NLMIXED procedure of the SAS system, is by determining

$$\mathrm{Cov}(\hat{\theta}_{0j}, \hat{\theta}_{1j}) = 1/2\left[\mathrm{Var}(\hat{\theta}_{0j} + \hat{\theta}_{1j}) - \mathrm{Var}(\hat{\theta}_{0j}) - \mathrm{Var}(\hat{\theta}_{1j})\right], \tag{5.5}$$

where $\mathrm{Var}(\theta_{0j} + \theta_{1j})$ is approximated with the delta method. We thus determined estimates of covariance parameters $(\hat{\theta}_{0j}, \hat{\theta}_{1j})$ as well as their asymptotic variance-covariance matrix $\mathrm{Cov}(\hat{\theta}_{0j}, \hat{\theta}_{1j})$.

The empirical Bayes estimates of the variance components are plugged into a mixed model analysis so that the contrasts of fixed effects are estimated based on $\hat{\theta}_j$. Linear contrasts of treatments or genotypes may be examined by means of a Wald test with the null hypothesis H$_0$: $L\binom{\beta}{u} = 0$, where L is a contrast vector. Note that contrary to the contrast vector in (3.10), the contrast vector in this chapter refers to fixed and random effects, which is indicated by a capital letter. If solely contrasts among fixed effects are considered, the part of L referring to the random effects is zero. The t-statistic is calculated by dividing the estimate of the linear contrast by the asymptotic variance of the estimate, which follows an approximate t-distribution:

$$\frac{L\left(\begin{array}{c}\hat{\beta}\\ \hat{u}\end{array}\right)}{\sqrt{L\hat{C}L'}} \sim t_\nu. \tag{5.6}$$

$\hat{C}$ denotes the approximate variance-covariance matrix of $\binom{\hat{\beta}}{\hat{u}}$ as obtained from the mixed model equations. As the variance estimates were computed using data of all spots, the degrees of freedom of a conventional mixed model are no longer valid. We therefore calculate the degrees of freedom ν of the t distribution with the Satterthwaite approximation according to McLean and Sanders (1988):

$$\nu \simeq 2(L'CL)^2/\mathrm{Var}(L'CL). \tag{5.7}$$

The degrees of freedom are calculated for each gene. In (5.7) and the following we omitted the index j to simplify notation and C is the approximate variance-covariance matrix of $(\hat{\beta} - \beta, \hat{u} - u)'$, given by the following equations:

$$\begin{aligned}C_{11} &= (X'V^{-1}X)^{+}\\ C_{12} &= -C_{11}X'V^{-1}ZG\\ C_{21} &= C'_{12}\\ C_{22} &= (Z'R^{-1}Z+G^{-1})^{-1}+GZ'V^{-1}XC_{11}X'V^{-1}ZG,\end{aligned}$$

where $G=\theta_1 I$ and $R=\theta_0 I$ are the variance-covariance matrices of the random effects and residual error, respectively, and $V=\text{Var}(y)=ZGZ'+R$. By using a first-order Taylor series expansion we get

$$L'CL \simeq (\hat{\theta}_0-\theta_0)\frac{\partial L'CL}{\partial\hat{\theta}_0}+(\hat{\theta}_1-\theta_1)\frac{\partial L'CL}{\partial\hat{\theta}_1}$$

and

$$\begin{aligned}\text{Var}(L'CL) \simeq \text{Var}(\hat{\theta}_0)\left(\frac{\partial L'CL}{\partial\hat{\theta}_0}\right)^2 &+ 2\text{Cov}(\hat{\theta}_0,\hat{\theta}_1)\frac{\partial L'CL}{\partial\hat{\theta}_0}\frac{\partial L'CL}{\partial\hat{\theta}_1}\\ &+\text{Var}(\hat{\theta}_1)\left(\frac{\partial L'CL}{\partial\hat{\theta}_1}\right)^2.\end{aligned}$$

As we consider contrasts among fixed effects, the derivatives reduce to

$$\frac{\partial L'CL}{\partial\theta_0}=K'C_{11}X'V^{-1}V^{-1}XC_{11}K$$

and

$$\frac{\partial L'CL}{\partial\theta_1}=K'C_{11}X'V^{-1}ZZ'V^{-1}XC_{11}K,$$

where K is the part of L that refers to the fixed effects. At this point, McLean and Sanders (1988) input the asymptotic covariance matrix for $\hat{\theta}_0$ and $\hat{\theta}_1$ (Giesbrecht, 1986) that is obtained through a restricted maximum likelihood analysis. However, we use the asymptotic variance-covariance matrix obtained by the empirical Bayes analysis.

5.2 Simulation study

The procedure described above was evaluated in a simulation study. We simulated 1000 datasets with 6000 spots each. The simulation model for a gene follows (5.1). For simplicity we assumed that only one fixed effect for genotype and one random effect for array exist. The illustrated method is useful when the arrays constitute incomplete blocks, i.e. when the number of treatments or genotypes to be compared exceeds two. Then we can benefit from the recovery of inter-array information. We therefore simulated 3 genotypes ($g_k, k = 1, 2, 3$) and investigated the new procedure for 3 different numbers of slides, namely 6, 9 and 12. Considering only array numbers that are multiples of three has the advantage that balanced incomplete block (BIB) designs are obtained. We thus have in each dataset $6000 \times 3 = 18000$ pairwise comparisons ($g_1 - g_2, g_2 - g_3, g_1 - g_3$). Data for 1500 spots was simulated with $g_1 = g_2 = g_3$, while the remaining 4500 spots were simulated so that $g_1 = g_2 \neq g_3$. As a result half of the comparisons followed the null hypothesis and the other half the alternative. The 9000 comparisons under the alternative were split into 10 groups; within each group the difference of genotype effects is the same. The 10 genotype differences were equally spaced, with the range depending on the number of arrays that were simulated. The random effects were taken from a bivariate lognormal distribution with mean μ and variance Σ, respectively. These were computed by exponentiating bivariate normal random variables with mean $\mu = \binom{-1.67}{-0.47}$ and dispersion matrix $\Sigma = \begin{pmatrix} 1.35 & 0.47 \\ 0.47 & 1.09 \end{pmatrix}$. These values were found when analysing the microarray experiment with maize endosperm described in Section 2.2 .

Within each simulated data set we determined pooled estimates for array and error variance. For each pairwise comparison we performed a t-test for differential expression between two genotypes. For comparison to the described procedure we performed an analysis by spot and calculated two kinds of t-tests. For the first test, the degrees of freedom were calculated according to the so-called containment method and for the second test they were approximated with the Satterthwaite method of ? (?). Both methods are implemented in the MIXED procedure of the SAS system (SAS Institute Inc., 1999). Power was assessed at nominal comparison-wise type I error rate of $\alpha = 0.05$.

Details for Implementation in SAS

The 'true' variance components for each spot were simulated using the MVN-macro (SAS Institute Inc., 2007). This macro creates multivariate normal random variables. Bivariate lognormal random variables were created by exponentiating a bivariate normal random variable.

The sums of squares were calculated using the GLM procedure. The Henderson type III sum of squares correspond to the type I sum of squares in SAS, when fixed effects are fit before random effects. Type I sum of squares were invoked via the 'SS1'-option in the MODEL statement. The coefficients of expected mean squares were read out of the ExpectedMeanSquares table using the SUBSTR-function. Degrees of freedom for array and error were given in the ModelANOVA- and OverallANOVA-tables, respectively.

The posterior distribution was modelled with the NLMIXED procedure. The distribution of the true variance components was indicated by the random statement. As the lognormal distribution is not supported, the logarithm of the variance components was specified as normally distributed. The scaled χ^2_k-distribution of the sum of squares was expressed in the MODEL statement using general(ll), where any log-likelihood function may be specified. We exploited the fact that the χ^2-distribution is a Gamma distribution with shape parameter $n = k/2$ and scale parameter two (Johnson, Kotz, & Balakrishnan, 1994, p. 450). To obtain the empirical Bayes estimates of the variance components, the PREDICT statement was used. The covariance between the variance components, which is needed to determine the Satterthwaite degrees of freedom, was calculated using a second PREDICT statement where the sum of the variance components was estimated.

The Satterthwaite degrees of freedom were calculated with IML. The rather cumbersome formula in McLean and Sanders (1988) reduces considerably when the contrast to be tested does not contain any random effects.

The data was then analysed in a spotwise manner with the MIXED procedure. The estimated variance components were fixed via the PARMS statement and the degrees of freedom calculated in IML were input through the 'DDF'-option in the MODEL statement.

5.3 Experimental data

As our variance estimating procedure performed well in simulations, we tested it with real data from the microarray described in Section 2.2. A linear mixed model according to (5.1) was assumed with fixed effects for genotype and dye and a random effect for slide. Empirical Bayes estimates of variance components were determined according to Section 5.1. Based on the model, dominance contrasts (i.e. hybrid minus parental mean) were tested, the null hypothesis being that the hybrid's expression equals that of the parental mean. The degrees of freedom were calculated with the Satterthwaite approximation. For comparative purposes, the same contrasts were tested with the same model and spotwise estimated variance components. The degrees of freedom were determined with the containment method and in another analysis with the method of Kenward and Roger. To adjust for multiplicity, the false discovery rate (Benjamini & Hochberg, 1995) was used and p-values below a cut-off of 0.05 were declared significant.

5.4 Results

Simulation Study

Table 5.1 shows the mean of the proportion of false positives over the 1000 simulated data sets, i.e. the proportion of non-differential contrasts that were declared to be significant. Contrary to the analysis of real data, for the simulation no multiplicity adjustment was performed for the simulation data, because methods for multiplicity adjustment such as procedures controlling the false discovery rate (FDR) require valid tests on a comparison-wise basis. For three and six arrays the test conducted with the pooled variance estimates was the only one that lay below the level α. For 12 arrays, the pooled variance test kept the level only for $\alpha = 0.001$. Here, the spotwise variance estimates with the degrees of freedom calculated by the containment method seem to perform a little better.

In Figure 5.1 the power of detecting true differences with the different tests is illustrated. The tests are demonstrated for designs with 6, 9 and 12 arrays and level $\alpha = 0.001$. On the abscissa we have the true differences d between genotypes and on the ordinate the proportion of signifi-

Table 5.1: *Observed comparison-wise type I error and standard error for different tests and array numbers*

number		$p < 0.01$		$p < 0.005$		$p < 0.001$	
of slides	method	p.f.p.	S.D.	p.f.p.	S.D.	p.f.p.	S.D.
	pooled	0.00719	0.00099	0.00312	0.00063	0.00040	0.00023
6	sS	0.01160	0.00126	0.00578	0.00090	0.00115	0.00042
	sC	0.01006	0.00117	0.00504	0.00083	0.00104	0.00040
	pooled	0.00957	0.00110	0.00459	0.00077	0.00079	0.00032
9	sS	0.01140	0.00121	0.00578	0.00085	0.00120	0.00040
	sC	0.01012	0.00114	0.00508	0.00079	0.00105	0.00039
	pooled	0.01024	0.00114	0.00507	0.00078	0.00097	0.00033
12	sS	0.01115	0.00118	0.00566	0.00084	0.00117	0.00039
	sC	0.01010	0.00113	0.00503	0.00078	0.00100	0.00036

p.f.p.=proportion of false positives; pooled=pooled variance estimates; sS=spotwise, dfs Satterthwaite; sC=spotwise, dfs Containment

cant tests among all tests performed for differences with true value d. For small differences between genotypes the detection power of the pooled variance test was lower than with the two spotwise methods. This may be explained by the fact that for the pooled variance test the level α is controlled in most cases, whereas it is exceeded with the other tests. For larger genotype differences, however, the detection power is considerably higher for the pooled variance test than with the two reference tests. This effect is most explicit for small array numbers.

Similar results were found by Wright and Simon (2003) when evaluating their random variance model. The superiority of the pooled residual variance estimates over spotwise estimates was smaller in experiments with higher array numbers. This confirms our assumption that pooled variance estimates are especially effective in experiments with a low number of replicates. The simulation showed that the procedure for the estimation of variance components proposed in Section 5.1 works well and is superior compared to spotwise estimation methods.

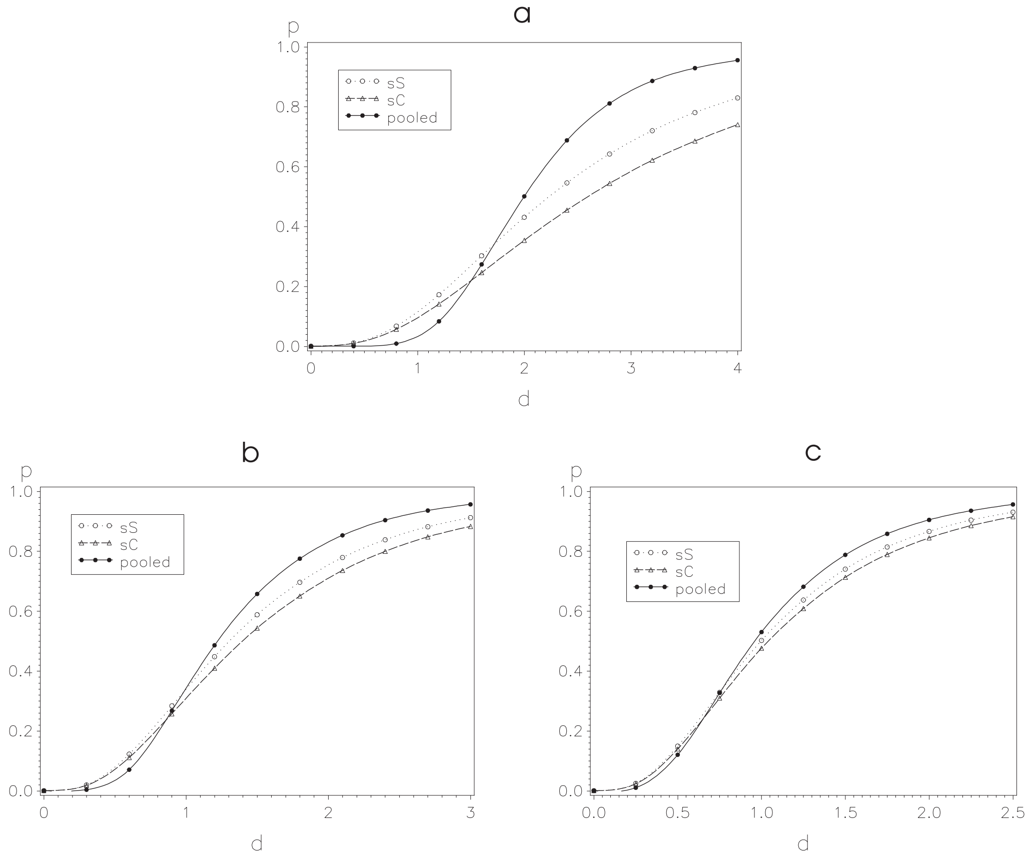


Figure 5.1: *Detection power for $p < 0.001$ level test by true mean difference d, (a) 6 arrays, (b) 9 arrays, (c) 12 arrays*

Experimental data

The log-transforms of the variance components are assumed to follow a normal distribution. The analysis gave parameter estimates for the mean $\hat{\mu} = \begin{pmatrix}-2.30\\-0.32\end{pmatrix}$ and dispersion matrix $\hat{\Sigma} = \begin{pmatrix}1.30 & 0.41\\0.41 & 0.67\end{pmatrix}$. To assess the validity of this assumption we simulated true variance components for each of the 6205 spots with these parameters. Based on these true variance components, sum of squares were simulated and order statistics of sum of squares were determined. This was repeated 100 times and the mean of the order statistics was plotted against the real sum of squares (Figure 5.2, top), with observed sum of squares as well as order statistics of simulated sum of squares log-transformed. At the upper and lower end the spots depart slightly from the bisecting line. Histograms of the observed sum of squares (Figure 5.2, bottom), however, reveal that only a minor proportion of the data lies in the tails of the distribution.

The subsequent analysis revealed that there was no great difference in

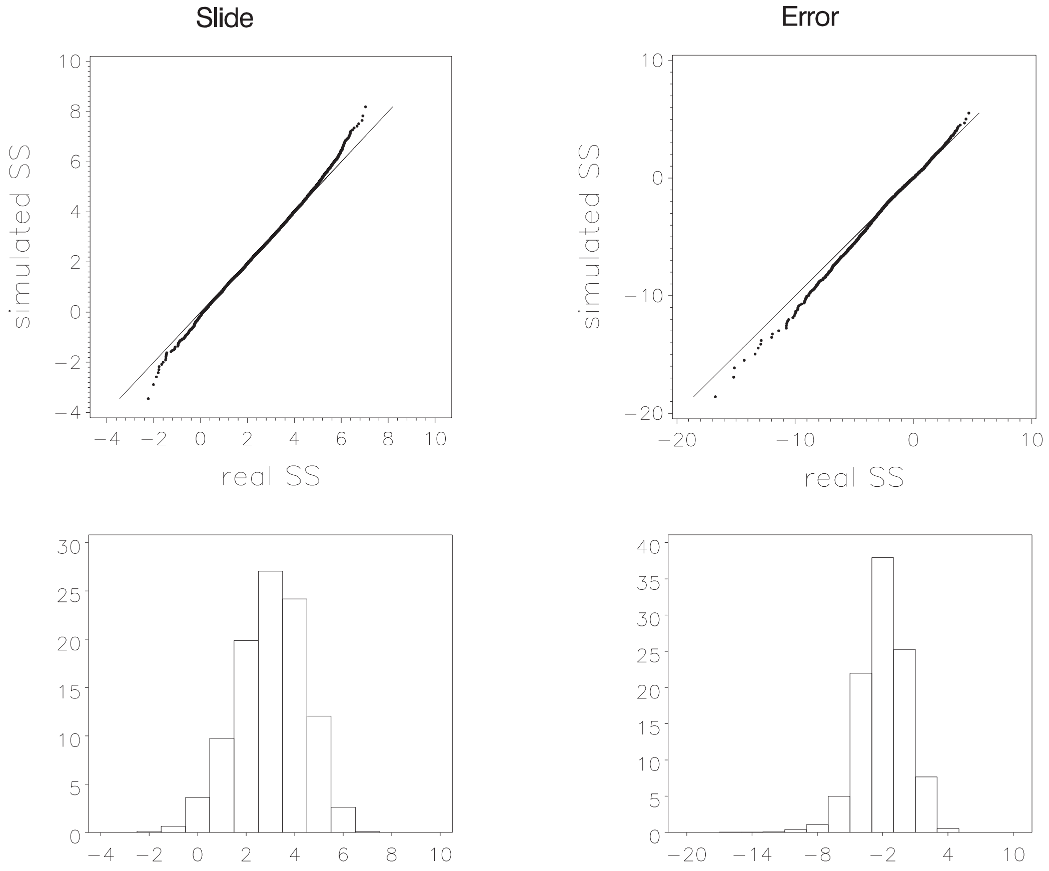


Figure 5.2: *Top: plots of order statistics for simulated sum of squares vs real sum of squares for slide and error. Bottom: histograms of real sum of squares for slide and error. Sum of squares in all graphs have been log-transformed*

the number of significant genes between different estimation methods regarding the contrast hybrid vs. parental mean. For one hybrid the pooled analysis gave slightly more significant contrasts than the other methods. For the second hybrid using pooled variance estimates resulted in slightly more significant genes than with spotwise estimates and degrees of freedom calculated with the containment method, but fewer significant genes than with the Satterthwaite method.

Table 5.2: *number of significant genes among a total of 6205 genes*

	pooled	spotwise	spotwise
		Satterthwaite	containment
UH250x301	174	170	136
UH301x250	194	232	185

5.5 Discussion

As microarray experiments are very cost-intensive, quite often the number of hybridisations is rather low and estimating variance components becomes a delicate task. Obviously, variance components may vary between genes and thus models with gene-specific variance components seem advisable. When using only data from one gene, however, the number of replicates may be low, resulting in imprecise estimates. We prepared an approach where the distribution of sum of squares given the true variance components is modeled. Based on this conditional distribution we calculated empirical Bayes estimates for the variance components of each spot. The method has the advantage that not only the residual array may be considered as random, but also the array variance. This may be of particular interest when contrasts between more than two genotypes are considered, e.g., when estimating dominance effects.

It was shown that the detection power of tests based on these pooled variance estimates is higher than the detection power of comparable tests, except for very small differences between genotypes. As scientists are more interested in noticeable differences than in marginal deviations, the lower power of the pooled variance test for small differences may be acceptable. Compared to the reference tests, the test based on pooled variance estimates in most cases keeps the level of the test, whereas the other tests rarely do.

In this study we focus on models with at least one further variance component beyond the residual variance, namely the array variance. The crucial feature of our methods is that it models the marginal distribution of ANOVA sums of squares rather than that of standard estimates of variance components (REML, ANOVA or other). This strategy was chosen to circumvent problems with non-positive variance estimates. ANOVA-estimates may become negative, and a prior distribution that includes negatives is undesirable. REML-estimates and ML-estimates have the deficiency that they may become zero, which complicates the fitting of a distribution for the true variance components. While distribution models with zeros exist (Aitchison, 1955), the conditional distribution of estimated variance components given the true variances is harder to derive. One possibility is to determine the conditional density of the REML-estimates given the true variance components for each of the following cases: (i) the REML-estimate of the array variance is larger than zero (ii) the REML-estimate of the array variance equals zero. The first two moments and other properties of the conditional distribution in each case may be determined by simulation. Furthermore a dummy variable is introduced that indicates whether case (i) or case (ii) holds true for each spot. The dummy variable follows the Bernoulli distribution with the parameter dependent on the true variance component. The likelihood for the REML-estimates of the variance components is obtained by integrating the conditional density multiplied with the corresponding parameter of the Bernoulli distribution. The conditional distribution of the REML-estimates, however, depends on the design of the experiment. Thus for each new experimental design the first and second moments of both distributions (array and residual variance) have to be simulated anew.

We chose the lognormal distribution as a prior for the variances. As shown in the previous section, the lognormal distribution fits quite well. Likewise, Lewin et al. (2006) and Cui et al. (2005) applied the lognormal distribution with good results Other distributions as a prior for the variances, however, are possible and seem to be worth future research. One possibility is the Johnson family of distributions (Johnson et al., 1994, p. 34), which contains the log-normal distribution as a special case. One of the transformations of a random variable X is $Z = \gamma + \delta \log(X - \xi), X \geq \xi$, the distribution of Z being unit normal. In Johnson (1949) the bivariate transformation family is described, which is particularly suited for modelling the joint distribution of array and residual variance components.

Analysis of the experimental data (Section 5.4) is completed by adjusting the p-values with the linear step-up procedure of Benjamini and Hochberg (1995). This procedure controls the false discovery rate under the assumption of independent test statistics. For the simulation study independence holds, but with real data this will be hardly the case due to co-expressing genes or spatial effects on the array. Benjamini and Yekutieli

(2001) show that the FDR is controlled for positively correlated one-sided test statistics. When estimating heterotic effects it is not only relevant to find genes that have an elevated expression level compared to the parents, but also to find genes where the hybrid is underexpressed with regard to the parents. Reiner-Benaim (2007) investigated the FDR with two-sided tests and different degrees of correlation and found the FDR to increase with increasing correlation. However, she could show that using the linear step-up procedure of Benjamini and Hochberg controls the FDR regardless of the proportion of true null hypotheses and dependence. Liu and Hwang (2007) propose a method to calculate the sample size while controlling the false discovery rate.

In this chapter models with two variance components were investigated. Sometimes, in addition to the effects related to hybridisation, effects regarding the design used to gain the biological material need to be taken into account (Chapter 3). If some of these are regarded as random, the described procedure may be extended. However, with more than two random effects and unbalanced models, the sums of squares for a mixed model are no longer necessarily independent and the exact form of the conditional joint distribution is more difficult to derive (Khatri, Krishnaiah, & Sen, 1977; Kotz, Balakrishnan, & Johnson, 2000). The type of approximation involved is similar to that for ANOVA F-tests under a mixed model, when the Satterthwaite method is used. According to Milliken and Johnson (1992, p. 252), the degrees of freedom for the tests may be approximated by the Satterthwaite method (notice that in Section 5.1 an extension by McLean and Sanders (1988) of the original Satterthwaite-approximation was used). The approximation in this case is twofold: the degrees of freedom are approximated and the linear combination of mean squares are not necessarily independently distributed as required by the approximation.

The problems caused by several random effects in the model does not occur with a fully Bayesian model as proposed by Gottardo et al. (2006). Their model detects differences in gene expression between two samples or genotypes. In the two-sample case, the effect of a sample on a specific gene is modeled as a random effect with a mixture of two singular normal distributions: the first corresponds to genes that are not differentially expressed, while the second component corresponds to differentially expressed genes. With more than two genotypes the model may also distinguish between different patterns of gene expression. When the number of samples grows larger, however, the model gets cumbersome and the number of parameters increases sharply, e.g., with three genotypes a mixture of five distributions is needed for the ability to distinguish the different expression patterns. This may be a problem when analysing large experiments as in Chapter 3. The Bayesian hierarchical model presented by Lewin et al. (2006) includes a differential effect between genotypes. To estimate heterosis contrasts, a suitable parametrization of the model would be necessary, e.g. by including effects for additivity and dominance.

The use of the Empirical Bayes estimates in a linear mixed model seems to be a competitive method when exploring heterosis. A further advantage is that with our approach the integrals may be approximated by Gaussian quadrature, as the logarithm of the random effects is assumed to be normal, whereas with the Bayes approaches posterior probabilities are calculated by Markov Chain Monte Carlo (MCMC) which is also comparatively time-consuming.

Chapter 6

General discussion

6.1 Importance and reliability of microarray data

The number of studies using microarrays to detect differentially expressed genes has exploded within the last ten years (Marshall, 2004). Microarrays provide a tool to gain information about a vast amount of genes by an experiment that is rather simple to conduct. The technology is popular not only in biology, but in medicine, too. In 2005 the Food and Drug Administration (FDA), the authority that is responsible for the admission of new pharmaceuticals in the U.S., approved the first microarray test, which is supposed to provide physicians with genetic information on their patients. This is assumed to further accelerate the rising tendency of microarray technology.

During the last years, concern emerged about the validity of microarray results. Performing the experiment there are lots of sources of variation, for example spatial effects on an array caused by dust particles or the dye effect. Regarding only the latter, already two sources of variation come into play: The two dyes, Cy3 and Cy5, are supposed to have different binding abilities. Furthermore, Cy5 is especially sensitive to atmospheric ozone levels, resulting in different experimental conditions for nice or dirty weather. Also there are technical differences between arrays of the different manufacturers. Tan et al. (2003) found that the set of genes differentially expressed in all of three investigated platforms is very low. Even when all these factors are neglected and only the raw data is regarded, slight changes in the statistical analyses may lead to largely contrasting results: Dave et al. (2004) found a correlation between survival length among patients with follicular lymphoma and molecular features of nonmalignant immune cells. During the analysis gene expression data was divided at random into training and test sets. Tibshirani (2005) performed the same analysis swapping training and test set, resulting in a non-significant correlation. Using Significance Analysis of Microarrays (SAM), a standard tool for the analysis of microarray data, Tibshirani (2005) likewise found no statistical evidence for a correlation between survival time and genetic features.

Otherwise, recent publications exist that paint a different picture: Irizarry, Warren, and Spencer (2004) report on a study where three platforms (spotted cDNA microarrays as well as oligonucleotide arrays) were tested in a total of ten laboratories. The precision of a platform in a laboratory was quantified by calculating correlations between different replicates. To assess accuracy, gene expression values of control genes were compared with the results of RT-PCR for the same genes. The authors found that the disagreement observed by other studies may be partly due to suboptimal statistical analysis. For instance, Kuo, Jenssen, Butte, Ohno-Machado, and Kohane (2002) and Tan et al. (2003) did not account for variability between laboratories in their studies. This laboratory effect is crucial and neglecting it may give poor results concerning reproducibility of results in different laboratories. Apart from the relatively large differences between laboratories using the same platform, Irizarry et al. (2004) found that the results from the best performing labs agreed rather well and that performance can be greatly improved when using alternative preprocessing and suitable statistical methods. An even more optimistic conclusion is drawn by the MicroArray Quality Control (MAQC) Consortium (Reid & Shi, 2006; Patterson et al., 2006). The MAQC project is initiated by FDA scientists to assess reproducibility, specificity, sensitivity and accuracy of microarray experiments. Seven microarray platforms and three alternative expression methodologies were investigated in different laboratories, showing intra platform consistency across test sites as well as a high level of inter platform concordance in terms of genes identified as differentially expressed. Within the SPP 'Heterosis in Plants' we also performed an evaluation of microarray reliability, which is described in Section 6.3.

6.2 The contribution of this thesis to methodology for microarray analyses and heterosis estimation

As stated by Irizarry et al. (2004) the use of adequate statistical methods is of utmost importance to obtain reliable information out of microarray data. This holds true for the standard case when two different tissue types are compared with regard to differential expression. When tackling more complex problems, even more emphasis should be placed on a proper analysis. In this thesis it was therefore attempted to improve statistical methods with regard to the investigation of the heterosis concept.

The estimation of heterosis differs from standard problems in that the expression of one genotype (the hybrid) is to be compared with the expression of the mean of two other genotypes (the parents). This may be adequately considered in the design of the experiment. If contrasts between any levels of factors are equally important, then it might be advisable to use a loop design. However, if the objective of the study concentrates on some specific contrasts that are far more important than others, this should be considered already in the design of the study. In this study, the contrast between a hybrid and the parental mean is regarded far more important as, e.g., the comparison between the two parents. As cDNA-microarrays may be hybridised with only two genotypes, more hybridisations should be performed with slides of hybrid-parent comparisons, as e.g., with parent-parent arrays. These arrays will be more informative with respect to heterosis. Likewise it is not very informative to hybridise two genotypes from different hybrid-parent pools together on one array, for instance hybrid UH002x301 with parent UH250. When keeping that in mind, an interesting contrast may be estimated with far more accuracy with an adapted design than with a standard microarray design.

Usually, microarray data show heterogeneous variance and data transformations are necessary before applying a linear model. The transformation the most commonly used is probably the log transformation, though other transformations are imaginable. One should be aware that transformations may have an impact on the results of a following statistical analysis. It was shown theoretically that a transformation of the linear predictor in a generalized linear model may remove partial heterosis, i.e. the hybrid is better than the parental mean, but not better than the better parent. Better parent heterosis cannot, however, be removed by transforming the linear predictor. It is shown exemplarily on phenotypic data that mid parent heterosis as well as better parent heterosis are not robust in terms of data transformations. This conclusion, although not entirely surprising, is all the more important as it can be not only transferred to microarray data, but is a general result concerning data transformations that may be important in other fields.

When analysing data aiming at the estimation of heterosis at least three genotypes are involved. With microarray data this means that arrays constitute incomplete blocks. It is therefore valuable to regard the array effect as random to benefit from the recovery of inter-array information. Analysing microarray data it is a common procedure to analyse data from each gene separately. Due to the relatively high costs of microarray hybridisations, few replicates are made and estimates of variance components are not too precise. Alternatively, variance components may be estimated across genes, having a large number of individual variance estimates disposable. Yet, with at least two random effects in the model, the procedure for the joint estimation of the variance components is not evident. When fitting a distribution to the sum of squares one avoids to fit a distribution to zeros, as could be the case when fitting the distribution to the variance component estimates. Assuming a bivariate lognormal distribution for the true variance components and fitting a distribution to the sum of squares, conditional on the variance components, results in estimates of variance components that are supposed to be more accurate compared to estimates based on the data of only one gene. When having more than two variance components in the model, the sum of squares might not be independent and the procedure given in Chapter 5 has to be expanded. As there are cases when a model with more than two variance components is reasonable, this is worthy of future research.

6.3 General results and experiences during my work within the SPP 'Heterosis in plants'

The statistical methods summarised in the preceding section wil hopefully increase the explanatory power and ameliorate the reproducibility of the results. Besides the development of statistical methods, my day-to-day business consisted to a large extent in the analysis of microarray and phenotypic data aiming at the exploration of the heterosis concept. Some results, experiences and findings of the collaborative work within the priority program 'Heterosis in plants' are summed up in the following paragraphs.

Our project partners from the University of Tübingen made the discovery that after scanning the microarray slides some highly expressed spots were saturated, because the scanner can only distinguish signals up to a certain intensity. When scanning the same slide at a lower intensity, dissatisfactory results were provided for lowly expressed spots (Figure 6.1 a). To get the best information for all spots, the array has to be scanned at different intensities.

We therefore developed a nonlinear latent regression model[1] (Piepho et al., 2006) that combines signals from multiple scans of one cDNA channel. The amount of effective expression product, which cannot be observed, is assumed to be the sum of an overall spot effect and an effect related to the scanning intensity: $\eta_{ij} = g_i + \alpha_j$, where η_{ij} is the latent value for the i-th

[1] Formulation of the model was done mainly by H.P. Piepho while my contribution is the implementation of the model in SAS software

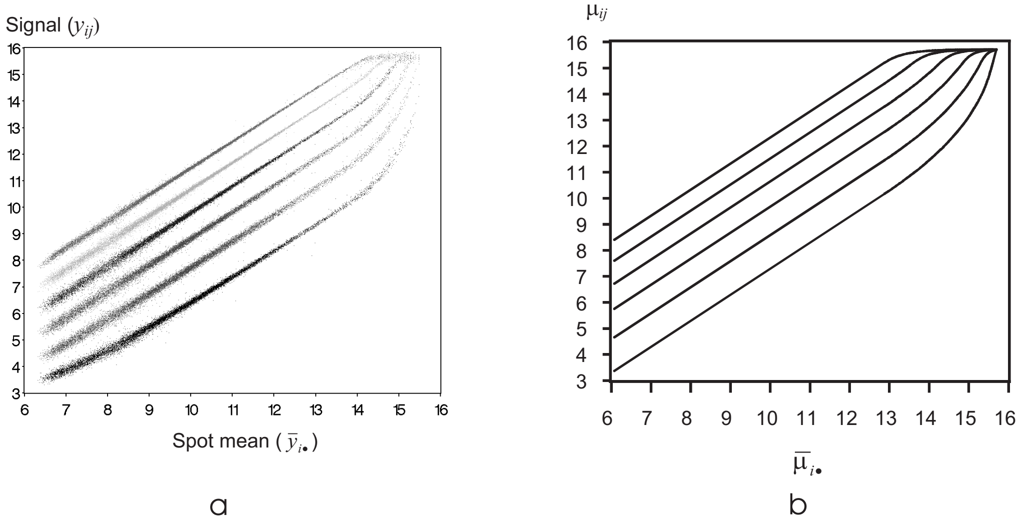


Figure 6.1: *(a) Plot of signals at six different scanning intensities vs. spot mean. Maize genotype UH005xUH301, University of Tübingen. (Piepho et al., 2006) (b) Plot of estimates of μ_{ij} vs. $\bar{\mu}_{i.}$ for different intensities*

spot at the j-th scanning intensity, g_i is the main effect of the i-th spot and α_j is the main effect of the j-th intensity. Below a certain threshold ϕ, the expected signal equals the effective expression product, while the model implies a nonlinear relationship between both variables when the effective expression product lies above ϕ:

$$\mu_{ij} = \begin{cases} \eta_{ij} & \text{if} \quad \eta_{ij} < \phi \\ \theta - \beta \exp(-\gamma\eta_{ij}) & \text{if} \quad \eta_{ij} \geq \phi. \end{cases}$$

Here, μ_{ij} is the expected signal for the i-th spot at the j-th scanning intensity, θ is the saturation limit, and β and γ are regression parameters (Figure 6.2). The observed signal y_{ij} is modelled as the expected signal plus an error term: $y_{ij} = \mu_{ij} + e_{ij}$. As scanning intensities seem to be more variable with low intensities, weights are computed for each spot as the inverse of the variance within a spot, which is predicted via loess regression. Intensity and spot effects as well as the parameters of the nonlinear function

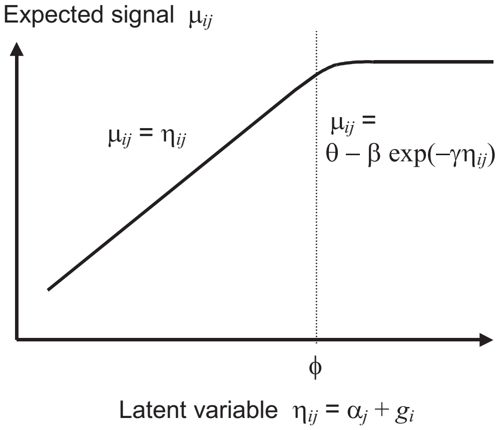

Expected signal μ_{ij}
$\mu_{ij} = \eta_{ij}$
$\mu_{ij} =$
$\theta - \beta \exp(-\gamma\eta_{ij})$
ϕ
Latent variable $\eta_{ij} = \alpha_j + g_i$

Figure 6.2: *Sketch of segmented model*

are estimated through an iterative algorithm: intensity effects and nonlinear function parameters are estimated alternately with the spot effects until the change in parameter estimates is smaller than some pre-defined limit. Spot effects are estimated separately for each spot by nonlinear least squares, while for the estimation of intensity effects and nonlinear function parameters by weighted nonlinear least squares the whole data set is used. This leads to a combined intensity value for each channel of a spot, where spots at the high end as well as spots at the low end of the intensity scale have reliable expression values. Figure 6.1 b, a plot of the expected signals for the different intensities vs. the mean of expected signals over all intensities, shows that the model fits very good.

The results of Irizarry et al. (2004) concerning reliability of microarray data coincide quite well with our findings. Within the priority program 'Heterosis in plants' several laboratories are performing microarray experiments and the wish to assess the reliability of the data emerged. A round robin experiment was conducted with three laboratories (University of

Tübingen, University of Hamburg, and University of Munich). Each laboratory performed four hybridisations of two maize genotypes (UH005, UH301xUH005), including two biological replicates and a dye swap. The analysis showed that the laboratory-effect was substantial. Despite this fact, two laboratories showed rather similar results. The data basis for the round robin experiment was considered not stable enough for further investigations, among other things due to failure of one hybridisation in one laboratory.

When analysing microarray data one is often asked to calculate fold changes, i.e. the ratio of estimated signal values between two samples. The foldchange may be a convenient and easily interpretable measure. It should not be forgotten, however, that it does not give any clue about the significance of a test result. Ranking of genes according to their interest for the biologist should therefore never be performed solely on the basis of fold changes. p-values of test statistics provide a much more reliable criteria. If only substantial differences between genotypes are of interest, then a cutoff for the foldchange may be used. Another widespread misconception is that the number of significant genes found in an analysis is an indicator for the quality thereof. This may easily be disproved by citing Klebanov, Qiu, Welle, and Yakovlev (2006):'...a given method is of limited utility if it consistently makes the same false discoveries from sample to sample.'

Some of the results obtained in my collaborative work (of a breeder's point of view) have already been published and will be described briefly. A comprehensive phenotypic investigation of heterosis in early maize roots was performed by our colleagues from the University of Tübingen

(Höcker et al., 2006)[2]. Several variables defining the root system (root length, root width, cortical cell length, number of seminal roots, lateral root density) of four inbred lines and twelve hybrids were collected. Mid-parent heterosis in the different traits was quantified via a linear model. The largest heterotic effects were found in the lateral root density, whereas root length between five to seven days after germination showed to be the most consistent heterotic trait. The analysis showed that heterosis is already manifesting during the very early stages of root development. The young maize root system was therefore subjected to a detailed molecular analysis of gene expression. A subsequent microarray analysis provided evidence that heterotic effects are also manifest on the gene level of young maize roots (results of microarray experiment are not yet published).

At the microarray experiment performed at the University of Munich (Uzarowska et al., 2006)[3] cDNA of meristem (tissue found in parts of the plant where growth takes place) of different maize inbred lines and hybrids was analysed. The data was normalized according to Section 2.3 and a mixed model was fitted. Besides the fixed effects for genotype, dye, and replicate and a random array effect, a seasonal effect for plant cultivation and an interaction between season and genotype was included, as some of the plants were cultivated in summer and some in winter. We tested for differential expression between the hybrids and the parental mean. A substantial part (38.1%) of the differentially expressed genes we found that have a known function, are associated with catalytic activities, whereas 33.3% are associated with binding activities. Furthermore, finding a high number of genes where the hybrid's expression level is higher than that of both parents supports the overdominance hypothesis, whereafter hetero-

[2] My contribution to the paper is the statistical analysis of the data.

[3] My contribution to the paper lies in the analysis of the microarray data

sis is due to an increased level of the heterozygote in certain responsible genes.

6.4 Concluding remark

The thesis has its field of application in plant breeding. Plant breeding has been practiced for centuries and the objectives are manifold. During the last century, agricultural focus lay mainly on the increase in yield. This was achieved e.g. by creating varieties that show high grain yields, that are more adapted to automated cultivation or that can better utilise fertilization. In effect, between 1961 and 1993 the worldwide cereal production rose from 877 million tons to 1894 million tons (Reeves, Pinstrup-Andersen, & Pandya-Lorch, 1999). According to Pingali (1999), as a result of the rapid population growth the global demand for rice, wheat and maize is expected to rise by 36, 40, and 47% between 1997 and 2020. Beyond the mere supply of food, agriculture today has to meet other challenges. Sustainable and environmentally friendly forms of agriculture are needed, and varieties that are more resistant against drought stress or better adapted to increased CO_2 levels could cope with future climate conditions.

A good understanding of the molecular basis of heterosis is vital for the breeding of such plants. This dissertation has made a contribution to the elucidation of the heterosis concept in improving statistical methods for microarray experiments. The issues addressed in this work were inspired by the idea of estimating heterosis, the methods presented might as well be useful when confronted with different objectives. Results may be referred to other areas and thus it is hoped that the dissertation will attract interest beyond the community of statisticians interested in plant breeding.

Appendix A

Monotonicity of the dominance ratio

The dominance ratio on the transformed scale is given by

$$\rho(\phi) = \frac{\delta(\phi)}{\alpha(\phi)} = \begin{cases} \frac{2\lambda_1^{\phi} - 1 - \lambda_1^{\phi}\lambda_2^{\phi}}{\lambda_1^{\phi}\lambda_2^{\phi} - 1} & \text{if } \phi \neq 0 \\ \frac{\ln\lambda_1 - \ln\lambda_2}{\ln\lambda_1 + \ln\lambda_2} & \text{if } \phi = 0 \end{cases} .$$

We show that $\rho(\phi)$ is monotonically decreasing for all ϕ.

PROOF.

As a result of l'Hospital's rule, $\lim_{\phi\to 0} \frac{2\lambda_1^{\phi} - 1 - \lambda_1^{\phi}\lambda_2^{\phi}}{\lambda_1^{\phi}\lambda_2^{\phi} - 1} = \frac{\ln\lambda_1 - \ln\lambda_2}{\ln\lambda_1 + \ln\lambda_2}$, $\rho(\phi)$ is a continuous function.

The first derivative of $\rho(\phi)$ with respect to ϕ is given by

$$\begin{aligned} \rho'(\phi) &= \frac{[2\lambda_1^{\phi}\ln\lambda_1 - \lambda_1^{\phi}\lambda_2^{\phi}\ln(\lambda_1\lambda_2)](\lambda_1^{\phi}\lambda_2^{\phi} - 1)}{(\lambda_1^{\phi}\lambda_2^{\phi} - 1)^2} \\ &\quad - \frac{[2\lambda_1^{\phi} - 1 - \lambda_1^{\phi}\lambda_2^{\phi}]\lambda_1^{\phi}\lambda_2^{\phi}\ln(\lambda_1\lambda_2)}{(\lambda_1^{\phi}\lambda_2^{\phi} - 1)^2} \\ &= \frac{2\lambda_1^{2\phi}\lambda_2^{\phi}\ln\lambda_1 - 2\lambda_1^{\phi}\ln\lambda_1 - \lambda_1^{2\phi}\lambda_2^{2\phi}\ln(\lambda_1\lambda_2) + \lambda_1^{\phi}\lambda_2^{\phi}\ln(\lambda_1\lambda_2)}{(\lambda_1^{\phi}\lambda_2^{\phi} - 1)^2} \\ &\quad - \frac{2\lambda_1^{2\phi}\lambda_2^{\phi}\ln(\lambda_1\lambda_2) - \lambda_1^{\phi}\lambda_2^{\phi}\ln(\lambda_1\lambda_2) - \lambda_1^{2\phi}\lambda_2^{2\phi}\ln(\lambda_1\lambda_2)}{(\lambda_1^{\phi}\lambda_2^{\phi} - 1)^2} \end{aligned}$$

$$\begin{aligned}
&= \frac{2[-\lambda_1^{2\phi}\lambda_2^{\phi}\ln\lambda_2 - \lambda_1^{\phi}\ln\lambda_1 + \lambda_1^{\phi}\lambda_2^{\phi}\ln(\lambda_1\lambda_2)]}{(\lambda_1^{\phi}\lambda_2^{\phi}-1)^2} \\
&= \frac{2}{(\lambda_1^{\phi}\lambda_2^{\phi}-1)^2}\left(-\lambda_1^{2\phi}\lambda_2^{\phi}\ln\lambda_2 - \lambda_1^{\phi}\ln\lambda_1 + \lambda_1^{\phi}\lambda_2^{\phi}\ln\lambda_1 + \lambda_1^{\phi}\lambda_2^{\phi}\ln\lambda_2\right) \\
&= \frac{2}{(\lambda_1^{\phi}\lambda_2^{\phi}-1)^2}\left[\lambda_1^{\phi}\lambda_2^{\phi}(1-\lambda_1^{\phi})\ln\lambda_2 + \lambda_1^{\phi}\lambda_2^{\phi}(1-\lambda_2^{-\phi})\ln\lambda_1\right] \\
&= \frac{2\lambda_1^{\phi}\lambda_2^{\phi}}{(\lambda_1^{\phi}\lambda_2^{\phi}-1)^2}\left[(1-\lambda_1^{\phi})\ln\lambda_2 + \left(1-\lambda_2^{-\phi}\right)\ln\lambda_1\right].
\end{aligned}$$

As the fraction on the left is positive for all ϕ, we show that

$$(1-\lambda_1^{\phi})\ln\lambda_2 + (1-\lambda_2^{-\phi})\ln\lambda_1 < 0 \quad \forall\ \phi \neq 0. \tag{A.1}$$

This is equivalent to

$$\begin{aligned}
-(\lambda_1^{\phi}-1)\ln\lambda_2 + (1-\lambda_2^{-\phi})\ln\lambda_1 &< 0 \\
\iff \qquad (1-\lambda_2^{-\phi})\ln\lambda_1 &< (\lambda_1^{\phi}-1)\ln\lambda_2 \\
\iff \qquad \tfrac{1-\lambda_2^{-\phi}}{\ln\lambda_2} &< \tfrac{\lambda_1^{\phi}-1}{\ln\lambda_1}
\end{aligned}$$

$\forall\ \phi \neq 0$. We define

$$h_1(\phi) = \frac{\lambda_1^{\phi}-1}{\ln\lambda_1} \qquad \text{and} \qquad h_2(\phi) = \frac{1-\lambda_2^{-\phi}}{\ln\lambda_2}.$$

The first and second derivatives of $h_1(\phi)$ and $h_2(\phi)$ are

$$h_1'(\phi) = \lambda_1^{\phi}, \qquad h_1''(\phi) = \lambda_1^{\phi}\ln\lambda_1,$$

and

$$h_2'(\phi) = \lambda_2^{-\phi}, \qquad h_2''(\phi) = -\lambda_2^{-\phi}\ln\lambda_2.$$

Then $h_1(0) = 0 = h_2(0)$ and $h_1'(0) = 1 = h_2'(0)$, i.e., the two functions touch in $\phi = 0$. Furthermore $h_1(\phi)$ is convex and $h_2(\phi)$ is concave as $h_1''(\phi) > 0$ and $h_2''(\phi) < 0$ for all ϕ. Therefore $h_2(\phi) < h_1(\phi)$ for all $\phi \neq 0$ and (A.1) holds true.

As $\rho'(\phi) < 0$ for all $\phi \neq 0$ and $\rho(\phi)$ is continuous for all ϕ, $\rho(\phi)$ is monotonically decreasing for all ϕ.

References

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, *50*, 901-908.

Angelis, L., Bora-Senta, E., & Moyssiadis, C. (2001). Optimal exact experimental designs with correlated errors through a simulated annealing algorithm. *Computational Statistics and Data Analysis*, *37*, 275-296.

Atkinson, A. C. (1985). *Plots, transformations, and regression*. London, UK: Clarendon Press.

Auger, D. L., Gray, A. D., Ream, T. S., Kato, A., Coe, E. H., & Birchler, J. A. (2005). Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics*, *169*, 389 - 397.

Baker, R. L., Nagda, S., Rodriguez-Zas, S. L., Southey, B. R., Audho, J. O., Aduda, E. O., et al. (2003). Resistance and resilience to gastrointestinal nematode parasites and relationships with productivity of red maasai, dorper and red maasai x dorper crossbred lambs in the sub-humid tropics. *Animal Science*, *76*, 119-136.

Becker, H. (1993). *Pflanzenzüchtung*. Stuttgart, D: Ulmer.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289 - 300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate under dependency. *Annals of Statistics*, *29*, 1165-1188.

Bernardo, R. (2002). *Breeding for quantitative traits in plants*. Woodbury, MN: Stemma Press.

Birchler, J. A., Auger, D. L., & Riddle, N. C. (2003). In search of the molecular basis of heterosis. *Plant Cell*, *15*, 2236 - 2239.

Boiteux, L. S., Hymand, J. R., Bach, I. C. B., Fonseca, M. E. N., Matthews, W. C., Roberts, P. A., et al. (2004). Employment of flanking codominant STS markers to estimate allelic substitution effects of a nematode resistance locus in carrot. *Euphytica*, *136*, 37-44.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society*, *B 26*, 211-246.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs*. New York, NY: John Wiley.

Connolly, J., & Wachendorf, M. (2001). Developing multisite dynamic models of mixed species plant communities. *Annals of Botany*, *88*, 703-712.

Cui, X., Hwang, J. T. G., & Qiu, J. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, *6*, 59-75.

Darwin, C. (1876). *The effects of cross and self fertilization in the vegetable kingdom*. New York, NY: Appleton.

Dave, S. S., Wright, G., Tan, B., A., R., Gascoyne, R. D., Chan, W. C., et al. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine*, *351*(21), 2159-2169.

Dobbin, K., & Simon, R. (2002). Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, *18*(11), 1438-1445.

Dudoit, S., Yang, Y. H., Speed, T. P., & Callow, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, *12*(1), 111-139.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., & Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, *18*(Suppl. 1), S105-S110.

East, E. M. (1908). Inbreeding in corn. *Report of the Connecticut Agricultural Experiment Station*, *1907*, 419-428.

Fai, A. H. T., & Cornelius, P. L. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation*, *54*, 363-378.

Falconer, D. S., & Mackay, T. F. C. (1996). *Quantitative genetics*. Essex, UK: Longman Group.

Freeman, G. H. (1976). On the selection of designs for comparative experiments. *Biometrics*, *32*, 195-199.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S., et al. (2004). Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics*, *167*(4), 1791 - 1799.

Giesbrecht, F. G. (1986). Analysis of data from incomplete block designs. *Biometrics*, *42*, 437-448.

Giesbrecht, F. G., & Burns, J. C. (1985). Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics*, *41*, 477-486.

Glover, F., & Laguna, M. (1997). *Tabu search*. Boston, MA: Kluwer Academic Publishers.

Gottardo, R., Raftery, A. E., Yeung, K. Y., & Bumgarner, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, *62*, 10-18.

Graw, J. (2006). *Genetik*. Berlin, D: Springer.

Guo, M., Rupe, M. A., Danilevskaya, O. N., Yang, X., & Hu, Z. (2003). Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. *Plant Journal*, *36*, 30-44.

Gurka, M. J., Edwards, L. J., Muller, K. E., & Kupper, L. L. (2006). Extending the box-cox transformation to the linear mixed model. *Journal of the Royal Statistical Society* A, *169*(2), 273-288.

Gurka, M. J., Edwards, L. J., & Nylander-French, L. (2007). Testing transformations for the linear mixed model. *Computational Statistics and Data Analysis*, *51*, 4297-4307.

Hetz, W., Hochholdinger, F., Schwall, M., & Feix, G. (1996). Isolation and characterisation of rtcs, a mutant deficient in the formation of nodal roots. *Plant Journal*, *10*, 845-857.

Höcker, N., Keller, B., Chollet, D., Descombes, P., Piepho, H. P., & Hochholdinger, F. (2007). High throughput qrt-pcr analyses of non-additive gene expression associated with heterosis manifestation in maize (*Zea mays* L.) primary roots reveal conserved trends in twelve hybrids (unpublished manuscript).

Höcker, N., Keller, B., Piepho, H. P., & Hochholdinger, F. (2006). Manifestation of heterosis during early maize (*Zea mays* L.) root development. *Theoretical and Applied Genetics*, *112*, 421-429.

Hsueh, H., Chen, J. J., & Kodell, R. L. (2003). Comparison of methods for estimating the number of true null hypotheses in multiple testing. *Journal of Biopharmaceutical Statistics*, *13*, 679-689.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and the quantification of differential expression. *Bioinformatics*, *18*(Suppl. 1), S96-S104.

Irizarry, R. A., Warren, D., & Spencer, F. (2004). Multiple-laboratory comparison of microarray platforms. *Nature Methods*, *2*(5), 345-349.

John, J. A., & Williams, E. R. (1995). *Cyclic and computer generated designs*. London, UK: Chapman & Hall.

Johnson, N. L. (1949). Bivariate distributions based on simple translation systems. *Biometrika*, *39*, 297-304.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions, Volume 1*. London, UK: John Wiley.

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, *79*, 853-862.

Keller, B., Emrich, K., Höcker, N., Hochholdinger, F., & Piepho, H.-P. (2005). Designing a microarray experiment to estimate dominance in maize (*Zea mays* L.). *Theoretical and Applied Genetics*, *111*, 57 - 64.

Keller, B., & Piepho, H.-P. (2005). Is heterosis an artefact governed by the choice of scale? *Euphytica*, *145*, 113-121.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983-997.

Kerr, M. K. (2003). Design considerations for efficient and effective microarray studies. *Biometrics*, *59*, 822-828.

Kerr, M. K., & Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, *2*, 183-201.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of the Institute of Statistical Mathematics*, *18*, 75-78.

Khatri, C. G., Krishnaiah, P. R., & Sen, P. (1977). A note on the joint distribution of correlated quadratic forms. *Journal of Statistical Planing and Inference*, *1*, 299-307.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671-680.

Klebanov, L., Qiu, X., Welle, ., S, & Yakovlev, A. (2006). Letter to the editor: Statistical methods and microarray data. *Nature Biotechnology*, *25*(1), 25-26.

Kollipara, K. P., Saab, I. N., Wych, R. D., Lauer, M. J., & Singletary, G. W. (2002). Expression profiling of reciprocal maize hybrids divergent for cold germination and desiccation tolerance. *Plant Physiology*, *129*, 974-992.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Multivariate distributions*. New York, NY: John Wiley.

Kuo, W., Jenssen, T., Butte, A., Ohno-Machado, L., & Kohane, I. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, *18*(3), 405-412.

Lamkey, K. R., & Edwards, J. W. (1998). Heterosis: theory and estimation. In *Proceedings of the 34th Illinois corn breeders' school* (p. 62-77). Urbana, IL.

Landgrebe, J., Bretz, F., & Brunner, E. (2006). Efficient design and analysis of two colour factorial microarray experiments. *Computational Statistics and Data Analysis*, *50*, 499-517.

Léon, J. (1994). Mating system and the effect of heterogeneity and heterozygosity on phenotypic stability. In *Biometrics in plant breeding: Applications of molecular markers* (p. 19-31). Wageningen, NL.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., & Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics*, *62*, 1-9.

Liu, P., & Hwang, J. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, *23*, 739-746.

Lönnstedt, I., & Speed, T. (2002). Replicated microarray data. *Statistica Sinica*, *12*, 31-46.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer.

Marshall, E. (2004). Getting the noise out of gene arrays. *Science*, *306*, 630-631.

Mather, K., & Jinks, J. L. (1977). *Introduction to biometrical genetics*. London, UK: Chapman & Hall.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London, UK: 2nd edition. Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: John Wiley.

McLean, R. A., & Sanders, W. L. (1988). Approximating degrees of freedom for standard errors in mixed linear models. *Proceedings of the Statistical Computing Section, American Statistical Association*, 50 -59.

Mead, R. (1988). *The design of experiments*. Cambridge, UK: Cambridge University Press.

Melchinger, A. E., Singh, M., Link, W., Utz, H. F., & Kittlitz, E. von. (1994). Heterosis and gene effects of multiplicative characters: Theoretical relationships and experimental results from *Vicia faba* L. *Theoretical and Applied Genetics*, *88*, 343-348.

Milliken, G. A., & Johnson, D. E. (1992). *Analysis of messy data. Volume 1: Designed experiments*. London, UK: Chapman & Hall.

Munson, P. (2001, Nov. 19). A 'consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. In *GeneLogic workshop on low level analysis of Affymetrix GeneChip data.*

Ni, N. Z., Sun, Q., Liu, Z., Wu, L., & Wang, X. (2000). Identification of a hybrid-specific expressed gene encoding novel RNA-binding protein in wheat seedling leaves using differential display of mRNA. *Molecular and General Genetics*, *263*, 934-938.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., et al. (2006). Performance comparison of one-color and two-color platforms within the microarray quality control (MAQC) project. *Nature Biotechnology*, *24*(9), 1140-1150.

Pearce, S. C. (1974). Optimality of design in plot experiments. *Proceedings of the 8th International Biometric Conference.*

Piepho, H. P. (1995). A simple procedure for yield component analysis. *Euphytica*, *84*, 43-48.

Piepho, H. P. (2005). Optimal allocation in designs for assessing heterosis from cDNA gene expression data. *Genetics*, *171*, 359-364.

Piepho, H. P., Büchse, A., & Emrich, K. (2003). A hitchhiker's guide to the mixed model analysis of randomized experiments. *Journal of Agronomy and Crop Science*, *189*, 310-322.

Piepho, H. P., & Emrich, K. (2005). Simultaneous confidence intervals for two estimable functions and their ratio under a linear model. *The American Statistician*, *59*, 292-300.

Piepho, H. P., Keller, B., Höcker, N., & Hochholdinger, F. (2006). Combining signals from spotted cDNA microarrays. *Bioinformatics*, *22*(7), 802-807.

Pingali, P. L. (1999). Role of heterosis in meeting world cereal demand in the 21st century. In J. G. Coors & S. Pandey (Eds.), *The genetics and exploitation of heterosis in crops* (p. 493-500). Madison, WI: ASA, CSSA, and SSSA.

Pinheiro, J. C., & 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.

Pukelsheim, F. (1993). *Optimal design theory*. New York, NY: John Wiley.

Reeves, T., Pinstrup-Andersen, P., & Pandya-Lorch, R. (1999). Food security and the role of agricultural research. In J. G. Coors & S. Pandey (Eds.), *The genetics and exploitation of heterosis in crops* (p. 1-5). Madison, WI: ASA, CSSA, and SSSA.

Reid, L. H., & Shi, L. (2006). The microarray quality control (MACQ) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, *24*(9), 1151-1161.

Reiner-Benaim, A. (2007). FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal*, *49*, 107-126.

Roumen, E. C. (1994). The inheritance of host-plant resistance and its effect on the relative infection efficiency of magnaporthe-grisea in rice cultivars. *Theoretical and Applied Genetics*, *89*, 498-503.

Sant, V. J., Patankar, A. G. P., Sarode, N. D., Mhase, L. B., Sainani, M. N., Deshmukh, R. B., et al. (1999). Potential of DNA markers in detecting divergence and in analysing heterosis in indian elite chickpea cultivars. *Theoretical and Applied Genetics*, *98*, 1217-1225.

SAS Institute Inc. (1999). *SAS/STAT® User's Guide, Version 8.* Cary, NC: SAS Institue Inc.

SAS Institute Inc. (1999/ 2002-2003). *SAS® software.* Cary, NC: SAS Institue Inc.

SAS Institute Inc. (2007). *Sample 509: Generate data from a multivari-*

ate normal distribution. http://support.sas.com/ctx/samples/index.jsp?sid=509&tab=details.

Schabenberger, O., & Pierce, F. J. (2002). *Contemporary statistical models for the plant and soil sciences*. Boca Raton, FL: CRC Press.

Schena, M. (2003). *Microarray analysis*. Hoboken, NJ: John Wiley.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. Hoboken, NJ: John Wiley.

Shull, G. F. (1908). The composition of a field of maize. *Report of the American Breeder's Association*, *5*, 51-59.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*, Issue 1.

Sparnaaij, L. D., & Bos, I. (1993). Component analysis of complex characters in plant breeding. *Euphytica*, *70*, 225-235.

Speed, T., & Yang, Y. H. (2002). Direct versus indirect design for cDNA microarray experiments. *Sankhya: The Indian Journal of Statistics*, *64*(A 3), 706-720.

Spilke, J., Hu, X., & Piepho, H. P. (2005). A simulation study on tests of hypotheses for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological and Environmental Statistics*, *10*, 374-389.

Stekel, D. (2003). *Microarray bioinformatics*. Cambridge, UK: Cambridge University Press.

Stuber, C. W. (1999). Biochemistry, molecular biology, and physiology of heterosis. In J. G. Coors & S. Pandey (Eds.), *Genetic and exploitation of heterosis in crops* (p. 31-48). Madison, WI: ASA, CSSA, and SSSA.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., et al. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, *31*(19), 5676-5684.

Tefera, H., & Peat, W. E. (1997). Genetics of grain yield and other agronomic characters in t'ef (eragrostis tef zucc trotter). *Euphytica*, *96*, 193-202.

Tibshirani, R. (2005). Immune signatures in follicular lymphoma: Letter to the editor. *New England Journal of Medicine*, *352*(14), 1496-1497.

Uzarowska, A., Keller, B., Piepho, H. P., Schwarz, G., Ingvardsen, C., Wenzel, G., et al. (2006). Comparative expression profiling in meristems of inbred-hybrid triplets of maize based on morphological investigations of heterosis for plant height. *Plant Molecular Biology*, *63*, 21-34.

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. New York, NY: Marcel Dekker, Inc.

Whitaker, D., Williams, E. R., & John, J. A. (2002). *CycDesigN: A package for the computer generation of experimental designs.* CSIRO,Canberra, AU.

Williams, E. R., & Talbot, M. (1993). *ALPHA+: Experimental designs for variety trials.* [Design User Manual]. CSIRO, Canberra and SASS, Edinburgh.

Wit, E., Nobile, A., & Khanin, R. (2005). Near optimal designs for dual channel microarray studies. *Journal of The Royal Statistical Society C*, *54*, 817-830.

Wright, G. W., & Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarrray experiments. *Bioinformatics*, *19*, 2448-2455.

Yang, X., Ye, K., & Hoeschele, I. (2002). Some E-optimal designs for cDNA microarray experiments. *ASA Proceedings of the Joint Statistical Meetings*, 3853-3954.

Yang, Y. H., Dudoit, S., Luu, P., & Speed, T. P. (2001, January). *Normalization for cDNA microarray data* (Tech. Rep. No. 589). Department of Statistics, University of California, Berkeley.