From the professorchip of Animal Breeding

Analyzing gene expression data with linear mixed models: Applications to variable pool sizes and biomarkers

Dissertation

for attainment of the academic degree
doctor of agriculturae

from the Faculty of Agricultural and Environmental Sciences
of the University of Rostock

Submitted by:

Henrik Rudolf
from Rostock

Rostock, July 2015

urn:nbn:de:gbv:28-diss2016-0062-1

Gutachter:

Manfred Schwerin, Prof. Dr. rer. nat.	Universität Rostock, Agrar- und Umweltwissenschaftliche Fakultät
Norbert Reinsch, Prof. Dr. agr.	Leibniz-Institut für Nutztierbiologie Institut für Genetik und Biometrie
Kaspar Bienefeld, Prof. Dr. agr.	Humboldt-Universität Berlin Lebenswissenschaftliche Fakultät

Datum der Verteidigung: 15.01.2016

To my Family.

Acknowledgements

At first I thank Norbert Reinsch for his many ingenious suggestions, valuable insights into research, and support. I have learnt very much from him.
I thank Manfred Schwerin for giving me the opportunity and provide the first report. For his willingness to discuss and support I sincerely thank Gerd Nuernberg. Many thanks go to the colleagues and staff of the Insitute for Genetics and Biometry at the Leibniz Institute for farm animal Biology who supported me during my time.
Without my family and Anna, it were not possible to write this thesis, I am very glad to have them.

This work has been funded by the German Federal Ministry of Research (BMBF project HyBee, PTJ 0315124D) and by the H.-Wilhelm-Schaumann-Stiftung and by the International Leibniz Graduate School on functional diversity in farm animals (ILGS DivA).

Contents

1 Introduction

Honeybees play an important role in nature as pollinators. At present, colonies of *Apis mellifera* are threatened by the mite *Varroa destructor*. The brood care of bees - consisting of certain behavioral traits, which might support resistance over the mite - is a complex social behavior and dependent on several circumstances. The root levels of brood care in the transcriptome can be investigated with gene expression profiling. This multi-step method allows for the measurement of thousands of genes simultaneously. A linear mixed model is an appropriate method for the analysis of measurements which are influenced by experimental factors and distortions. It offers the possibility of considering systematic sources of variation and to adjust the question of interest with respect to other experimental factors.

The choice of design in a microarray experiment depends on the experimental question, number of individuals to analyze, number of available arrays, the choice of one- or two-color platforms, and also the financial background. With DNA microarrays, the mRNA of an individual or pooled sample is isolated and transcribed into cDNA and then the fluorescent emission of labelled complementary hybridization is scanned. Pooling - the blending of individual mRNA samples - reduces hybridization costs and compensates for insufficient amounts of mRNA. For a given number of arrays the inclusion of more individuals via pooling also has an advantage when testing the hypotheses, as for differentially expressed transcripts theoretically get a higher power. On the other hand, due to blending and the logarithmic transformation, a part of the normalization for statistical analysis of gene expression data, biases can be introduced. However, by adequately choosing the design and statis-

tical analysis method, their impact can be excluded from contrasts and hypotheses tests.

The blending of individual mRNA samples into a pool also has consequences on the variance of measurements. Furthermore, an additional technical error can occur due to pooling. In the authors opinion, a corresponding variance component called blending error variance has to be included in a model for data from pooled sample designs. Four experimental data sets were examined in order to study the relevance of the blending error variance.

The variance structure of the data is also of importance, if multivariate methods come into play, which is often the case in a biomarker search. The goal in such an analysis is the selection of a set of genes whose joint expression pattern is a basis for a rule to classify unknown samples. The preparation of the data to a suitable gene expression matrix with one expression vector per sample is generally not straightforward. This is the case, if a design with replications is applied. In experiments with two-color arrays, the entries of the gene expression matrix have to be further extracted out of differences of expression values. This transformation is essential for the ability to find good candidate biomarker genes.

In the background of a honeybee project about the hygienic behavior against the parasitic mite *V. destructor*, methods presented in this thesis were initially developed. But they are applicable to a much wider class of experiments. For the choice of design in the case of variable pool sizes (the number of individuals per pool), a condition for unbiased contrasts was derived. The relevance of a technical error due to the blending of individual samples into pools was investigated with four experimental data sets of different species and in several simulations. For the transformation of two-color array data to a gene expression matrix suitable for a biomarker search the known back-transformation to single channel expression led to poor classification performance. A method aimed at the best possible preservation of the variance structure was developed.

This introductory chapter is concluded as follows. Because the main methodol-

ogy used in this thesis for tackling the aforementioned problems were linear mixed models, the next section contains some related background with basic explanations about contrast and variance functions. Then a brief description of the gene expression experiment from the German Federal Mnistry of Education and Research (BMBF)-project HyBee is presented. Thereafter, the issues picked up in the above mentioned areas of research will be described, to be discussed in the three main chapters of this thesis.

1.1 A few basics to linear mixed models

A linear mixed model allows for fixed and random effects. It is used here to model the vector $\mathbf{y}$ of logarithmically transformed expression for a transcript via design matrices $\mathbf{X}$ for fixed effects and $\mathbf{Z}$ for random effects.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1.1}$$

where $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, E(\mathbf{u}) = 0, E(\mathbf{e}) = 0, V(\mathbf{u}) = \mathbf{G}, V(\mathbf{e}) = \mathbf{I}_n\sigma_e^2, Cov(\mathbf{u}, \mathbf{e}) = 0$. In this work, the covariance matrix of random effects $\mathbf{G}$ consists of one block per random effect in $\mathbf{u}$. This is called the variance component form of the general model. The variance structure modeled for $\mathbf{y}$ in equation 1.1 is $\mathbf{V}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^\top + \mathbf{I}_n\sigma_e^2$, where $\mathbf{I}_n$ denotes the $(n \times n)$ identity matrix. Generally, with such a variance structure heterogeneity and correlations can be considered. There are some ways to fit a linear mixed model, for instance, the restricted maximum likelihood (REML) method. It estimates the residual variance and each of the variance components, and then the estimates of the fixed and random effects are obtained by solving the mixed model equations.

To compare estimated parameters $\hat{\boldsymbol{\beta}}$ a contrast might be used. $(l \times p)$-matrix $\mathbf{K}^\top = (k_{ij})_{i=1\ldots l, j=1\ldots p}$, where p is the number of parameters and l the dimension, is called contrast matrix if $\sum_j k_{ij} = 0, \forall i$. A contrast function is estimable if and only

if $\mathbf{K}^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^-(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}) = \mathbf{K}^\top$, where $(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^-$ is a generalized inverse of $\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}$. This applies to models where the design matrix $\mathbf{X}$ is not of full rank. With a contrast $\mathbf{K}^\top$ and estimated $\boldsymbol{\beta}$ hypotheses $\mathbf{K}^\top\hat{\boldsymbol{\beta}} = 0$ can be tested. The corresponding Wald-statistic $\frac{1}{l}(\mathbf{K}^\top\hat{\boldsymbol{\beta}})^\top(\mathbf{K}^\top(\mathbf{X}^\top\hat{\mathbf{V}}^{-1}\mathbf{X})^-\mathbf{K})^{-1}(\mathbf{K}^\top\hat{\boldsymbol{\beta}})$ has an F-distribution. For details and a complete overview of linear model theory, the reader is referred to the work of Searle (1971).

With a (one-dimensional) contrast, it can then be tested whether a gene or transcript is differentially expressed between (two) treatments. Contrast functions can also be used for comparing two designs, e.g. regarding the power to detect differences between treatments. The calculation of variance functions $\mathbf{K}^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^-\mathbf{K}$, using the design matrices $\mathbf{X}$ and assuming known $\mathbf{V}$, yields a smaller variance function for the contrast for a 'better' design.

1.2 Background to honeybee and the HyBee project

Colonies of *A. mellifera* are threatened by the parasitic mite *V. destructor*. The mites reproduce inside the brood cells. Hygienic behavior is considered an effective control strategy against brood diseases. The afore-mentioned project (BMBF project HyBee, PTJ 0315124D) was one part of the FUGATO-plus research series with the title: HyBee - developing molecular methods for the selection of pathogen resistant honeybees based on gene expression differences of hygienic and not hygienic worker bees. Hygienic behavior was analyzed at the individual level of worker bees. Various experiments with regard to this rare individual behavior for instance in breeding and olfactory learning were executed. The gene expression experiments which investigated the two behavioral classes of worker bees from behavioral assays, are a subject and motivation for this thesis. The first part was a preliminary experiment which included pooled samples with 22 arrays. The second part was conducted in the final year of the project with 64 arrays per tissue and single samples.

The basis for the evaluation of worker bees were infrared video observations of in-

dividually marked worker bees showing their activities in a hive with a section of *Varroa*-parasitized brood. The detected hygienic bees were then compared in a microarray experiment against control bees and between different colonies. The objective for the development of the gene expression experimental design was to use all hygienic bees, whose identification took great efforts, with a given number of arrays. Several bee colonies participated in the behavioral assay of 48 hours duration, which was repeated seven times. After each assay the bees were freeze-killed with liquid nitrogen to secure age-specific expression. For the gene expression experiment, it was decided to compare bees with different hygienic status only inside the same colony and the same assay. Thus, the blocking factors for the experiment were assay and colony. The number of available arrays in the project was limited and beforehand the number of hygienic bees was unknown. Therefore it was planned to pool bees detected as hygienic from the same block and oppose a pool of controls, which were available in large numbers, to be measured with a two-color array. This principle leads to variable pool sizes.

1.3 Design principles for pooling experiments

There are many possible ways to combine pooled samples on different arrays. For economical reasons the number of arrays is usually limited, and here it is assumed to be fixed. In a study like the bee project, where in the beginning the number of extraordinary individuals within a large number of animals is unknown, an effective use of resources might be to consider pools of variable size. This would include all individuals with a rare property like the hygienic behavior of honeybees. It was considered a problem, that biases are introduced due to pooling (Kendziorski et al., 2005) and with a variable pool size, the variance of measurements is no longer homogeneous. As a consequence it was recommended to avoid variable pool sizes in microarray experiments.

The root of the problem is that pooling happens on the original scale of mRNA sam-

ples of individuals and the statistical analysis takes place after data normalization, including a logarithmic transformation. In Zhang et al. (2007) the relationships of expectation and variance between pooled samples and individual samples were theoretically derived. The situation that the bias and the variance of measurements of pooled samples on the logarithmic scale both depend on the pool size, motivated the search for possibilities to correct for flexible pooling (experimental designs with pooled samples allowing a different number of individual samples to be blended into a pool). First, intuitively one might guess, the size of the control pool shall equal the treatment pool to secure estimates of main treatment effects, which can be adjusted for pool size. As in experiments like the bee project, the number of controls is usually large, to each pool of the 'special' individuals an equally sized pool of controls can easily be opposed. Secondly, it has to be investigated how the variance heterogeneity can be handled in the analysis.

How this principle basically works is shown in the following for a minimal example. We consider one factor of interest with two levels: treatment (with two pool sizes) and control, to be measured with four one-color arrays. Such an experiment can be modeled with a linear model $\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}^*$ is the data vector, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta} = [\ \mu \quad \beta_1 \quad \beta_2\ ]^\top$ is the parameter vector and $\mathbf{e}$ the residual term. Here, measurements assumed to be independent and therewith $\mathbf{V} = \mathbf{V}(\mathbf{y}^*)$ is diagonal. Furthermore, it is assumed that the vector $\mathbf{h}$ contains the biases due to pooling of the data vector $\mathbf{y}^*$. The entries of $\mathbf{h}$ and the covariance matrix $\mathbf{V}$ depend on the pool size and thus get a different index for each pool size. Assigning to the first treatment pool a control pool of the same size and also for the second treatment pool, the entries in $\mathbf{h}$ are pairwise equal, as well as elements on the diagonal of $\mathbf{V}$. A generalized least squares estimate of unbiased measurements $\mathbf{y}$ would be $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}$. For the biased measurements $\mathbf{y}^* = \mathbf{y} + \mathbf{h}$ the estimate of $\boldsymbol{\beta}$ is also biased and denoted by $\boldsymbol{\beta}^*$. Evaluating the influence of biased measurements on the contrast of main effects, $\mathbf{K}^\top = [\ 0 \quad 1 \quad -1\ ]$ is applied on $\boldsymbol{\beta}^*$. Then we

can write:

$$\begin{aligned}\mathbf{K}^\top\boldsymbol{\beta}^* &= \mathbf{K}^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{\text{-}}\mathbf{X}^\top\mathbf{V}^{\text{-}1}(\mathbf{y}+\mathbf{h}) \\ &= \mathbf{K}^\top\hat{\boldsymbol{\beta}} + \mathbf{K}^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{\text{-}}\mathbf{X}^\top\mathbf{V}^{\text{-}1}\mathbf{h}. \end{aligned} \tag{1.2}$$

For unbiasedness the right term in equation (1.2) has to be checked, first calculating:

$$\begin{aligned}&(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{\text{-}}\mathbf{X}^\top\mathbf{V}^{\text{-}1}\mathbf{h} \\ &= \left(\begin{bmatrix} \frac{1}{v_1} & \frac{1}{v_1} & \frac{1}{v_2} & \frac{1}{v_2} \\ \frac{1}{v_1} & 0 & \frac{1}{v_2} & 0 \\ 0 & \frac{1}{v_1} & 0 & \frac{1}{v_2} \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}\right)^{\text{-}}\begin{bmatrix} \frac{1}{v_1} & \frac{1}{v_1} & \frac{1}{v_2} & \frac{1}{v_2} \\ \frac{1}{v_1} & 0 & \frac{1}{v_2} & 0 \\ 0 & \frac{1}{v_1} & 0 & \frac{1}{v_2} \end{bmatrix}\begin{bmatrix} h_1 \\ h_1 \\ h_2 \\ h_2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{2}{v_1}+\frac{2}{v_2} & \frac{1}{v_1}+\frac{1}{v_2} & \frac{1}{v_1}+\frac{1}{v_2} \\ \frac{1}{v_1}+\frac{1}{v_2} & \frac{1}{v_1}+\frac{1}{v_2} & 0 \\ \frac{1}{v_1}+\frac{1}{v_2} & 0 & \frac{1}{v_1}+\frac{1}{v_2} \end{bmatrix}^{\text{-}}\begin{bmatrix} \frac{2h_1}{v_1}+\frac{2h_2}{v_2} \\ \frac{h_1}{v_1}+\frac{h_2}{v_2} \\ \frac{h_1}{v_1}+\frac{h_2}{v_2} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{v_1v_2}{v_1+v_2} & 0 \\ 0 & 0 & \frac{v_1v_2}{v_1+v_2} \end{bmatrix}\begin{bmatrix} \frac{2h_1}{v_1}+\frac{2h_2}{v_2} \\ \frac{h_1}{v_1}+\frac{h_2}{v_2} \\ \frac{h_1}{v_1}+\frac{h_2}{v_2} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{h_1v_2+h_2v_1}{v_1+v_2} \\ \frac{h_1v_2+h_2v_1}{v_1+v_2} \end{bmatrix}. \end{aligned} \tag{1.3}$$

From the finding of equation 1.3 it follows that $\mathbf{K}^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{\text{-}}\mathbf{X}^\top\mathbf{V}^{\text{-}1}\mathbf{h} = 0$. Hence, the difference of main effects of the two levels is unbiased, if for each pool of the treatment an equally sized pool of controls is opposed.

In chapter 2, it is shown that balancing pool sizes between treatment groups also suffices in the general case of more treatments and more complex variance structures. Together with a general pooled sample design with flexible pooling, the way of statistical analysis also had to be refined. Therefore, a model for analysis of pooling experiments was developed. It allows us to adequately handle the bias caused by pooling and the variance heterogeneity, which were the main concerns about microarray experiments, where the pool size is not fixed.

1.4 Estimation and tests of blending error variance

Biases reflect the impact of pooling on the expectation, while the influence of the technical error due to pooling on the variance of pooled samples and on differential expression was underestimated. Analyzing experiments with variable pool size makes it possible to estimate the technical error due to pooling. It appears through unequal shares of individual mRNA in a pool. In Figure 1.1, it is shown how blending

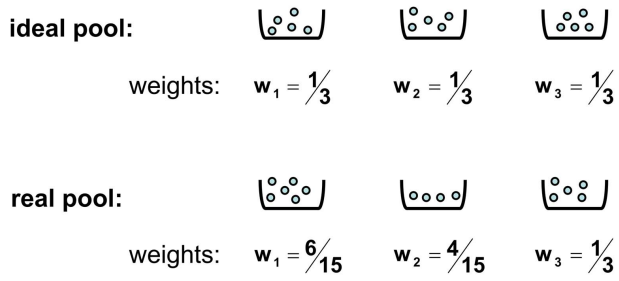


Figure 1.1: Scheme for blending of individual samples at DNA level

of individual samples might be imagined at the level of aliquots. For a pool blended from three individual samples, the fictive weights in an ideal pool and a real pool are displayed. It is important to remark that unequal shares might arise differently for every transcript, instead of assuming overall weights, which correspond to the volumetric amount of the individual samples in a pool.

In Zhang et al. (2007), the pooling technical variance was assigned to the variance of weights of individuals in a pool with expectation $\frac{1}{\gamma}$, for a pool size γ. This concept was further developed and extended to the blending error variance corresponding to distortions of the expression level of pools at the logarithmic scale, where it can be estimated as a random effect in a mixed model.

In chapter 3, four experimental data sets were investigated. The blending error variance was estimated by using an EM-REML script developed for the case of two random effects or rather three variance components in a linear mixed model. Transcript by transcript was tested, if blending error variance was significantly different from zero, by a REML log likelihood ratio test of the hypothesis $\sigma_2^2 = 0$. Correction for multiple testing according to a false discovery rate of 5% was applied. Further-

more, basics of estimation of variance components with REML and the simulation of pooled sample gene expression data with respect to the log-transformation are given.

1.5 Searching RNA-marker for hygienic behavior

The data from the gene expression experiment with 64 arrays from the tissue of the mushroom body was chosen for the biomarker search for hygienic behavior, because recognition of olfactory sensitivity was attributed to that area of the bee brain. A biomarker search is done for selecting a set of genes, whose joint gene expression pattern allows the derivation of a rule for the classification of unknown samples to classes.

To start the search algorithm with gene expression data (64 arrays), certain preparations are necessary, including setting up a gene expression matrix with one data vector per sample (1,...,95). Then learning takes place with multivariate methods. Applying a scheme which re-samples training and test set, a classifier is calculated and then evaluated at the test set. Repetition of this procedure, many thousand times, yields a forecast of the misclassification rate (prediction error) expected with independent or validation data (e.g. Dziuda, 2010). The two-color array design from HyBee included replicated use of bees and therefore correlated measurements have been generated. A good preservation of the variance structure of the data is essential for biomarker search. Only then can a classifier be expected to be generalizable and allow the classification of unknown samples to pre-defined classes. In the beginning a simple approach deriving expression vectors for individual samples from the differences in log expression via back-transformation prepended several learning methods. The performance in this setting was poor, independent of the chosen method, with prediction errors of around 40% (results not shown).

In chapter 4, a new approach is presented which includes an estimation of random bee effects, as a summary of all phenotypical information. In this way, the correla-

tions between measurements via the covariance matrix of the mixed model and also the multiple usage of bees are considered, although both masked in differences.

Further structure and related publications

Altogether, selecting the optimal design for the special circumstances of a study and properly modeling and analyzing microarray data are challenges which this work deals with, with emphasis on pooling and biomarker search.
The research topics introduced above yield the following structure: Using the matrices corresponding to the design of a pooling experiment, an unbiasedness condition for contrasts is given in chapter 2. Together with other aspects of flexible pooling, it was published in the journal *Statistical Applications in Genetics and Molecular Biology*. The empirical investigations into the inaccuracies of the blending of individual samples into pools from the chapter 3 were included in a paper Rudolf *et al.* (2015), which was accepted for publication in *BMC Genomics*. A manuscript including the presented findings of chapter 4 is in preparation (Rudolf *et al.*, 2016). Presented results, in relation to other work and consequences for future research are summarized and generally discussed in chapter 5. A summary in German is also given in chapter 6. The Appendix gives some supplementary information to chapter 3, an R-script for estimation in linear mixed models with EM-REML, the result of generalized linear mixed model analyses to estimate intra-class correlations of traits from the behavioral assay of HyBee, as well as calculations of variance functions for minimal examples of designs with regard to the power to detect differentially expressed genes.

2 Flexible pooling in gene expression profiles: design and statistical modeling of experiments for unbiased contrasts

2.1 Introduction

Pooling is a widely used method in gene expression profiling, which is employed to control the costs of the arrays or to deal with an insufficient amount of RNA (see e.g. Kerr (2003)). At the same time pooling reduces the biological variation and the variance of contrasts, what favors statistical testing. Several researchers investigated pooling with regard to power of tests for finding differentially expressed genes. Kendziorski *et al.* (2003) showed that for the precision of estimates pooling is most advantageous when biological variability is larger than technical. Most of the work in finding optimal designs related to pooling is done for a fixed number of individuals and varying number of arrays according to a uniform pool size. The assumption of a fixed number of allocated arrays is economically more reasonable. Peng *et al.* (2003) concluded that optimal pooling designs can be found to meet statistical requirements while minimizing total cost. To make data convenient for statistical methods, e.g. linear models, the log-transformation became an accepted

standard. That pooling happens on the original scale makes modeling more complex. Zhang *et al.* (2007) assessed the difference of pool-signals and their averaged contributions on the log-scale. Further, they mentioned that changing sizes of pools in an experiment should be avoided for two main reasons. First, it leads to biases in measurements depending on the size of pools and second also to variance heterogeneity. However, situations occur in which mRNA-pooling with varying numbers of individuals per pool is favored, for example if comparisons between subjects in a family of animals are performed, where it is impossible to quantify the size of the subgroups in advance. One aim is to include more individuals than in the uniform pool size design, if the number of arrays is limited. We were motivated by contemplating the design for a honey bee experiment, where rare cases - workers with a certain kind of behavior, see Thakur *et al.* (1997) - ought to be compared with controls from the same colony. In an elaborate experiment a certain number of cases was identified, too much to be compared individually. Therefore pooling RNA of cases from the same colony was considered. This led to variable pool sizes, as the number of identified cases per colony was not uniform. We asked the question of experimental designs allowing for unbiased contrasts and tests of hypotheses by choosing the number and size of control pools adequately. There are similar situations in plants and animal families with a high number of offspring like in fish, mice or pigs with an abundant number of individuals, which can be taken as controls. Also the investigation of human diseases may lead to experiments where few cases are found among many potential controls, e.g. Jacobsen *et al.* (2007).

In this article, we study microarray experiments with varying pool sizes (flexible pooling) for an efficient use of individuals and address the drawbacks given above. Under certain restrictions we show how to eliminate the bias caused by pooling for correctly testing hypotheses. In this context, we derive a balancing condition for suitable designs which allows us to estimate contrasts without bias. Our data are usually not only heteroscedastic but may, in principle, also be correlated, due to repeated mixtures and replications. Hence, the two random effects introduced

to handle the variance heterogeneity also reflect the impact of the mixtures. Our approach with different pool sizes increases the possibilities of designing pooled microarray experiments. The results for one-color arrays are extended to two-color arrays. Finally, consequences on experimental power are examined by means of simulation.

2.2 Theory

2.2.1 Modeling bias on the log-scale

In this section we consider the statistical model for mRNA pooling according to Zhang *et al.* (2007) and review their results regarding the bias and variance of the pooled samples. This is the starting point for our new approach to unbalanced sampling by means of flexible pool sizes.

To enforce normality in the data, a log transformation is deployed (see Geller *et al.* (2003)). In a general setting with T treatments and N individuals per treatment we describe the logarithm of the true gene expression level for each subject by

$$\log m_{tq} = \mu_t + \epsilon_{tq},$$

where μ_t is the fixed treatment effect for a gene, ϵ_{tq} is the biological error modeled by independent random variables which are identically distributed as $N(0, \sigma_b^2)$, $t = 1, \ldots, T$ and $q = 1, \ldots, N$.

Pooling is now applied for controlling costs, securing enough mRNA for hybridization or reducing variance of observations. Only in an ideal pool each individual has an equally sized contribution. In general, the true gene expression level of one gene in a pool of size γ is built from convex linear combinations of γ elements like

$$m_{tj}^p = \sum_{k=1}^{\gamma} \left(w_{tjk} \cdot m_{tjk} \right),$$

where m_{tjk} is the true gene expression level of individual k in pool j of treatment t with $t = 1, \ldots, T$, $j = 1, \ldots, S$ and S the number of pools per treatment. The weights are built from $N(1, \sigma_z^2)$- distributed random variables z by $w_{tjk} = z_{tjk} / \sum_{k=1}^{\gamma} z_{tjk}$ and have expectation $1/\gamma$ and variance σ_w^2. As pooling happens on the original scale the distribution of log-pool-signals is analytically intractable. Like in Zhang *et al.* (2007) we assume that convex linear combinations of log-normal distributed random variables can be approximated by log-normal distributions. Their computations, based on a Taylor Expansion and the Delta Method, lead to approximative formulas for expectation and variance of pool-signals on the log-scale. Supposing the distribution of $\log\left(m_{tj}^p\right)$ can be adequately approximated by a normal distribution, the mean log gene expression levels μ_t^p of the pools are:

$$\mu_t^p = \mathrm{E}\left[\log\left(m_{tj}^p\right)\right] \approx \mu_t + \frac{\sigma_b^2}{2} - \frac{1}{2\gamma}\left(e^{\sigma_b^2} - 1\right)\left(1 + \frac{\gamma - 1}{\gamma}\sigma_z^2\right) \tag{2.1}$$

and the biological variance of the log of the expression level of a pool can be expressed by

$$\sigma_b^{2,p} \sim \frac{1}{\gamma}\left(e^{\sigma_b^2} - 1\right)\left(1 + \gamma^2\sigma_w^2\right). \tag{2.2}$$

Here we mention that in formula (2.2) the coefficient of σ_w^2 has to be γ^2, which was incorrectly displayed in the original paper of Zhang *et al.* (2007) (details can be found in the Appendix).

Until now we assumed a balanced model, with the same number of pools per treatment S, number of individuals per pool γ and number of arrays. The disadvantage of a uniform pool size experiment is that any random fluctuation in the number of individuals forces the experimenters either to buy more arrays or leave out individuals. As in the already mentioned honey bee experiment, situations may occur when varying numbers of subjects have to be included in the gene expression analysis without changing the number of pools or arrays. In doing so we can use more individuals as in the uniform pool size experiments and expect to increase power in tests for differentially expressed genes (DEGs). We combine pools of different size in

one experiment and do this in such a way, that the pools assigned to each level of a treatment agree both in number and sequence of their sizes. This is necessary for a bias-free analysis, as we will see in the next sections. In our setting A is the number of different pool sizes for each of the T treatments, P_i the number of pools of size γ_i in each treatment and n is the total number of pools $n = T\sum_{i=1}^{A} P_i = TS$. For example, in a microarray experiment with only two treatments we oppose to each pool of the first treatment group an equally sized pool of controls. In this minimal case we apply formula (2.1) for $\mu_t^p - \mu_t$ to two treatments ($T = 2$) and to one pool per treatment ($P_1 = 1$). For the next pairs of pools we check again the size and either need to increment P_1 or switch to another pool size. An application of this design is presented in the Appendix.

Formula (2.1) can be utilized for defining the biases

$$h_i = \frac{\sigma_b^2}{2} - \frac{1}{2\gamma_i}\left(e^{\sigma_b^2} - 1\right)\left(1 + \frac{\gamma_i - 1}{\gamma_i}\sigma_z^2\right), \tag{2.3}$$

for pools of size γ_i, $i = 1, \ldots, A$. They depend on the biological variance σ_b^2, the pooling technical variance σ_z^2 and the pool size γ_i for the pool i. Due to our flexible pooling experimental design we see in the following that the contrasts defined by the practical question of testing for differences in gene expression between treatments are unbiased.

2.2.2 Conditions for unbiased hypothesis testing

The logarithm of the observed gene expression levels is modeled in the pooled setting as

$$y_{tj}^p = \log\left(m_{tj}^p\right) + e_{tj}, \tag{2.4}$$

where e_{tj} is the technical error which is assumed to be independent, identically distributed as $N(0, \sigma_t^2)$, $t = 1, \ldots, T$ and $j = 1, \ldots, S$. The matrix reformulation of

the previous equality can be written as

$$\mathbf{y}^p = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \mathbf{e}, \tag{2.5}$$

where $\mathbf{y}^p = [y_{11}, y_{21}, \ldots, y_{T1}, y_{12} \ldots, y_{T2}, \ldots, y_{TS}]^\top$ is the data vector with covariance matrix $\mathbf{V}$, $\mathbf{X} = \mathbf{1}_S \otimes \begin{bmatrix} \mathbf{1}_T & \mathbf{I}_T \end{bmatrix}$ is the design matrix, $\mathbf{h} = [h_1\mathbf{1}_{TP_1}, \cdots, h_A\mathbf{1}_{TP_A}]^\top$ is the vector of nonnegative biases, $\mathbf{e} = [e_{11}, \ldots, e_{TS}]^\top$ is the vector of technical errors. Here we denote the Kronecker product by $\otimes$ and vectors of length n built from n elements a by $a\mathbf{1}_n = [a, ..., a]^\top$ and the $n \times n$ identity matrix as $\mathbf{I}_n$. The fixed effect $\boldsymbol{\beta} = [\mu, \beta_1, \ldots, \beta_T]$ includes an intercept μ and the treatment effects $\beta_t = \mu_t - \mu$. We consider a contrast $\mathbf{K}^\top$, which fulfills the necessary and sufficient condition for estimability $\mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X} = \mathbf{K}^\top$, where $\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^-$ is the generalized inverse of $\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}$.

For the generalized least squares estimator $\hat{\boldsymbol{\beta}}$ we can write

$$\begin{aligned}
\mathrm{E}\left(\mathbf{K}^\top\hat{\boldsymbol{\beta}}\right) &= \mathrm{E}\left(\mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}^p\right) \\
&= \mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\mathrm{E}\left(\mathbf{y}^p\right) \\
&= \mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{h}\right] \\
&= \mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{h} \\
&= \mathbf{K}^\top\boldsymbol{\beta} + \mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{h}
\end{aligned} \tag{2.6}$$

Hence, $\mathbf{K}^\top\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\mathbf{K}^\top\boldsymbol{\beta}$ if and only if

$$\mathbf{K}^\top \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{h} = 0. \tag{2.7}$$

In order to test for differences of gene expression levels between treatments we choose the contrast function

$$\mathbf{K}^\top = \left[\mathbf{0}_{T-1}\ \mathbf{1}_{T-1}\ \text{-}\mathbf{I}_{T-1}\right]. \tag{2.8}$$

As already mentioned, we consider designs which have an equal number of pools of

the same size for each treatment group and for each possible pool size. In consequence, the vector of biases $\mathbf{h}$ contains only elements of equal values for each pool of the same size. We show now that the bias of the chosen contrast can be removed from the analysis by these suitable designs. Each individual occurs in one pool only, so that we can assume to have independent pools. In this heteroscedastic model we have a diagonal structure of $\mathbf{V} = \mathbf{\Phi} \otimes \mathbf{I}_T$, with $\mathbf{\Phi} = \begin{bmatrix} v_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & v_S \end{bmatrix}$, where variances of independent pools are on the diagonal with their values reflecting pool sizes. At first we evaluate:

$$\begin{aligned}
\left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^- &= \left[\left(\mathbf{1}_S \otimes \begin{bmatrix} \mathbf{1}_T & \mathbf{I}_T \end{bmatrix} \right)^\top \left([\mathbf{\Phi} \otimes \mathbf{I}_T] \right)^{-1} \left(\mathbf{1}_S \otimes \begin{bmatrix} \mathbf{1}_T & \mathbf{I}_T \end{bmatrix} \right) \right]^- \\
&= \left(\mathbf{1}_S^\top \mathbf{\Phi}^{-1} \mathbf{1}_S \otimes \begin{bmatrix} T & 1 & \cdots & 1 \\ 1 & & & \\ \vdots & & \mathbf{I}_T & \\ 1 & & & \end{bmatrix} \right)^- \\
&= \frac{1}{\sum_{j=1}^{S} \frac{1}{v_j}} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \mathbf{I}_T & \\ 0 & & \end{bmatrix}.
\end{aligned}$$

For $\tilde{\mathbf{h}} = [h_1 \mathbf{1}_{P_1}, \ldots, h_A \mathbf{1}_{P_A}]^\top$ we have $\mathbf{h} = \tilde{\mathbf{h}} \otimes \mathbf{1}_T$ and

$$\begin{aligned}
\mathbf{X}^\top \mathbf{V}^{-1} \tilde{\mathbf{h}} \otimes \mathbf{1}_T &= \left(\mathbf{1}_S \otimes \begin{bmatrix} \mathbf{1}_T & \mathbf{I}_T \end{bmatrix} \right)^\top \left([\mathbf{\Phi} \otimes \mathbf{I}_T] \right)^{-1} \tilde{\mathbf{h}} \otimes \mathbf{1}_T \\
&= \mathbf{1}_S^\top \otimes \begin{bmatrix} \mathbf{1}_T^\top \\ \mathbf{I}_T \end{bmatrix} \left(\mathbf{\Phi}^{-1} \tilde{\mathbf{h}} \otimes \mathbf{1}_T \right) \\
&= \mathbf{1}_S^\top \mathbf{\Phi}^{-1} \tilde{\mathbf{h}} \otimes \begin{bmatrix} T \\ \mathbf{1}_T \end{bmatrix} = \sum_{j=1}^{S} \frac{h_j}{v_j} \begin{bmatrix} T \\ \mathbf{1}_T \end{bmatrix}.
\end{aligned}$$

Hence, we get

$$\mathbf{K}^\top\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{h} = [\mathbf{0}_{T-1}\ \mathbf{1}_{T-1}\ \text{-}\mathbf{I}_{T-1}]\ \frac{1}{\sum_{j=1}^{S}\frac{1}{v_j}}\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \mathbf{I}_T & \\ 0 & & \end{bmatrix}\sum_{j=1}^{S}\frac{h_j}{v_j}\begin{bmatrix} T \\ \mathbf{1}_T \end{bmatrix}$$

$$= \frac{\sum_{j=1}^{S}\frac{h_j}{v_j}}{\sum_{j=1}^{S}\frac{1}{v_j}}\,[\mathbf{0}_{T-1}\ \mathbf{1}_{T-1}\ \text{-}\mathbf{I}_{T-1}]\begin{bmatrix} 0 \\ \mathbf{1}_T \end{bmatrix} = \mathbf{0}_{T-1}.$$

This means that the difference between the treatments can be unbiasedly estimated, although the estimations of the fixed effects show a systematic error due to variable pool sizes.

We generalize our statistical model to the case of d multivariate measurements taken from the same individuals. For instance, one may consider the same quantity of interest at different times (longitudinal data) or measurements which are distinguished by a second factor (e.g. tissue) at one time. In this case the entries of the data vector $\mathbf{y}^*$ and the bias vector $\mathbf{h}^*$ have two indexes according to the number of pools ($n = ST$) and the d levels of the second factor. Hence, the vectors $\mathbf{y}^*$ of measured log-pool-intensities and $\mathbf{h}^*$ of biases can be written as:

$\mathbf{y}^* = [y_{11}, \ldots, y_{1d}, y_{21}, \ldots, y_{2d}, \ldots, y_{n1}, \ldots, y_{nd}]^\top$,

$\mathbf{h}^* = [h_{11}, \ldots, h_{1d}, h_{21}, \ldots, h_{2d}, \ldots, h_{n1}, \ldots, h_{nd}]^\top$.

Note that the new design matrix $\mathbf{X}^* = \mathbf{X} \otimes \mathbf{I}_d$ has the dimension $STd \times STd$, and the bias vector is $\mathbf{h}^* = \mathbf{h} \otimes \mathbf{1}_d$. Moreover, a similar structure expressed by means of the matrix $\mathbf{I}_d$ is found for the contrast asserting the question of testing for the differences in gene expression levels between treatments $\mathbf{K}^{\top^*} = \mathbf{K}^\top \otimes \mathbf{I}_d$. $\mathbf{V}^*$ has a block-diagonal structure and two consecutive blocks of $\mathbf{V}^*$ have the same elements. The covariance matrix $\mathbf{V}^*$ consists of repeated $(d \times d)$-blocks with entries derived from the variances of the univariate case and a fundamental correlation matrix $\mathbf{W}$. For the matrix $\mathbf{W}$ various correlation structures can be imagined, for instance autoregressive dependencies for measurements over time or just correlations between

different tissues. Thus the n $(d \times d)$-blocks can be merged in $\mathbf{V}^* = \mathbf{V} \otimes \mathbf{W}$. Beforehand we saw that condition (2.7) ensures the unbiasedness of the contrast. For a design with two factors this equality is illustrated by an example in the Appendix. We now check the relation (2.7) in the multivariate setting

$$\mathbf{K}^{\top *} \left(\mathbf{X}^{*\top} \mathbf{V}^{*^{-1}} \mathbf{X}^* \right)^{-} \mathbf{X}^{*\top} \mathbf{V}^{*^{-1}} \mathbf{h}^* = \mathbf{0}.$$

This can be written as

$$\mathbf{K}^{\top} \otimes \mathbf{I}_d \left[(\mathbf{X} \otimes \mathbf{I}_d)^{\top} (\mathbf{V} \otimes \mathbf{W})^{-1} (\mathbf{X} \otimes \mathbf{I}_d) \right]^{-} (\mathbf{X} \otimes \mathbf{I}_d)^{\top} (\mathbf{V} \otimes \mathbf{W})^{-1} \mathbf{h}^* = \mathbf{0},$$

which is equivalent to

$$\mathbf{K}^{\top} \otimes \mathbf{I}_d \left[\mathbf{X}^{\top} \mathbf{V}^{-1} \mathbf{X} \otimes \mathbf{W}^{-1} \right]^{-} \left(\mathbf{X}^{\top} \mathbf{V}^{-1} \otimes \mathbf{W}^{-1} \right) \mathbf{h} \otimes \mathbf{1}_d = \mathbf{0}.$$

Hence, it holds that

$$\mathbf{K}^{\top} \left(\mathbf{X}^{\top} \mathbf{V}^{-1} \mathbf{X} \right)^{-} \mathbf{X}^{\top} \mathbf{V}^{-1} \mathbf{h} \otimes \mathbf{1}_d = \mathbf{0}. \tag{2.9}$$

This means that the unbiasedness for the univariate case is valid for the multivariate one as well. In the second case the maintenance of the condition $\mathbf{K}^{*\top} \mathbf{h}^* = \mathbf{0}$ in the design requires more effort, since we need to level the pool sizes for each additional factor, too.

Summary of rules for designs with flexible pooling The overall principle derived from the above calculations is to secure the balance pool sizes between treatment groups. This symmetry in the structure of the pool sizes across treatments ensures that the condition $\mathbf{K}^{\top} \mathbf{h} = \mathbf{0}$ holds. Hence, the pooling biases do not matter anymore, since they are eliminated by others with a corresponding value in the contrast function. In the particular case of one treatment and a control we need pairs of

equally sized pools, the first pool of each pair from the treatment and the second one from the control. Further each pair is required to share the same combination of other possible explanatory variables. In this case the unbiasedness condition is fulfilled.

2.2.3 Generalized covariance structure

Through the appropriate choice of the design the bias is eliminated. For a correct analysis of experiments with flexible pool sizes we have to deal with the appearing inhomogeneity in the variance of the measurements. We use a mixed model approach, mainly because the estimation of variance components allows reasonable testing, comparison of designs and yields a better error structure. Several authors (e.g. Wolfinger *et al.* (2001), Tempelman (2005)) review microarray experiments with mixed effects models. As discussed in Rosa *et al.* (2005) these models are preferable because in a fixed effects model the error term refers only to the lowest level of replication, leading to overestimation of power as well to inflation of the type I error. In microarray experiments varieties consisting of the mRNA samples and arrays are considered as blocks. In our model the random effects are shaped by repeated mixtures, replications and pool sizes, which leads to biologically interpretable variance components.

Decomposition of pool variance

Starting from the approximation of the biological variance of a pool of size γ_i as given in the relation (2.2), we can estimate the total variance of this pool as

$$\begin{aligned} v_i &\approx \frac{1}{\gamma_i}\left(e^{\sigma_b^2}-1\right)\left(1+\gamma_i^2\sigma_w^2\right)+\sigma_t^2 \\ &= \left(e^{\sigma_b^2}-1\right)\frac{1}{\gamma_i}+\left(e^{\sigma_b^2}-1\right)\left(\sigma_w^2\gamma_i\right)+\sigma_t^2 \\ &\approx \left(e^{\sigma_b^2}-1\right)\frac{1}{\gamma_i}+\left(e^{\sigma_b^2}-1\right)\sigma_z^2\frac{\gamma_i-1}{\gamma_i^2}+\sigma_t^2. \end{aligned} \tag{2.10}$$

We elaborate three variance components to be estimated with a mixed model, whose interpretation allows a new approach to pooled microarray experiments. We define $\sigma_1^2 := e^{\sigma_b^2} - 1$ and $\sigma_2^2 := (e^{\sigma_b^2} - 1)\sigma_z^2$. Biological variance between pools is described by σ_1^2 and the variance of mixtures by σ_2^2. The technical variance σ_t^2, caused by hybridization and measurement, corresponds to the array errors. The possible replication steps are illustrated in the Appendix in Figure 2.4. If there are more measurements than pools, we get replications of pools and the question what is replicated becomes important. If only the 'measurement steps' like reverse transcription, labeling and hybridization are repeated, we deal with the random pool effects and the residual variance, which contains the technical error. Otherwise, if some steps like extraction, mixing or amplification are executed, additional sources of variation appear, which we describe by a second random effect for every new mixture of a pool.

Mixed model and random effects

Our aim is now to formulate an appropriate mixed model for the observed pooled gene expression levels and to estimate its variance components. Therefore, in our model equation we introduce two more random effects for pools of different sizes and their repeated mixtures, and also an error term for their replications. We denote by L_i the cardinality of the set of different mixtures of the pool (t, i) for gene expression measurements. We have tuples $\left\{\left(w_{tik}^l\right)_{k=1,\ldots,\gamma_i, l=1,\ldots,L_i}\right\}$ of weights for each mixture of the pool (t, i). Then the true gene expression level in the pool (t, i) and for the mixture l is $m_{til}^p = \sum_{k=1}^{\gamma_i} w_{tik}^l m_{tik}$. Additionally, R_{il} is the number of the technical replications of the mixture (t, i, l) leading us to the observed values $y_{tilr}^p = \log(m_{til}^p) + e_{tilr}$. Then the model contains a fixed treatment effect β_t, a vector of biases $\mathbf{h}$, a random effect for pools $\mathbf{u}^1$, a nested effect for repeated mixtures $\mathbf{u}^2$, and the technical error

$$y_{tilr}^p = \mu + \beta_t + h_i + u_{ti}^1 + u_{til}^2 + e_{tilr}, \tag{2.11}$$

where $u_{ti}^1 \sim N\left(0, \frac{1}{\gamma_i}\sigma_1^2\right)$, $u_{til}^2 \sim N\left(0, \frac{\gamma_i - 1}{\gamma_i^2}\sigma_2^2\right)$, $e_{tilr} \sim N\left(0, \sigma_t^2\right)$, $t = 1, \ldots, T$, $i = 1, \ldots, A$, $l = 1, \ldots, L_i$ and $r = 1, \ldots, R_{il}$.

The data is correlated due to repeated mixtures or replications of the pools of the same size for each treatment. After grouping pools of the same size together we can write the covariance matrix of the observations from treatment t as

$$\mathrm{Cov}\left(\mathbf{y}_{t\ldots}^p, \mathbf{y}_{t\ldots}^p\right) = \bigoplus_{i=1}^{A} \Sigma_i,$$

where $\Sigma_i = \frac{1}{\gamma_i}\sigma_1^2 \mathbf{1}_{P_i \times P_i} + \bigoplus_{i=1}^{L_i} \mathbf{M}_{il}$, $\mathbf{M}_{il} = \frac{\gamma_i - 1}{\gamma_i^2}\sigma_2^2 \mathbf{1}_{R_{iL_i} \times R_{iL_i}} + \sigma_t^2 \mathbf{I}_{R_{iL_i}}$ and $P_i = \sum_{l=1}^{L_i} R_{il}$. Writing the vector $\mathbf{y}^p$ by pool sizes and alternating the blocks of the same sizes across treatments we obtain its covariance matrix as $\mathbf{V} = \bigoplus_{i=1}^{A} (\mathbf{1}_T \otimes \Sigma_i)$. This is needed for the extension of the validity of condition (2.7) to the case of generalized covariance between measurements.

Our bias vector is concatenated out of sequences of h_i corresponding to different pools of size γ_i: $\mathbf{h} = \underset{i=1}{\overset{A}{\mathbf{conc}}} (\mathbf{1}_T \otimes h_i P_i)$ and the design matrix $\mathbf{X}$ equals $\underset{i=1}{\overset{A}{\mathbf{conc}}} \{[\mathbf{1}_T \mathbf{I}_T] \otimes \mathbf{1}_{P_i}\}$. We check now condition (2.7) for our mixed model.

We calculate first

$$\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} = \left(\sum_{i=1}^{A} \mathbf{1}_{P_i}^\top \Sigma_i^{-1} \mathbf{1}_{P_i} \right) \begin{bmatrix} T & \mathbf{1}_T^\top \\ \mathbf{1}_T & \mathbf{I}_T \end{bmatrix},$$

and

$$\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{h} = \left(\sum_{i=1}^{A} h_i \mathbf{1}_{P_i}^\top \Sigma_i^{-1} \mathbf{1}_{P_i} \right) \begin{bmatrix} T \\ \mathbf{1}_T \end{bmatrix}.$$

We get:

$$\mathbf{K}^\top \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^{-} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{h} = \mathbf{K}^\top \left(\sum_{i=1}^{A} \mathbf{1}_{P_i}^\top \Sigma_i^{-1} \mathbf{1}_{P_i} \right)^{-1} \begin{bmatrix} 0 & \mathbf{0}_T^\top \\ \mathbf{0}_T & \mathbf{I}_T \end{bmatrix} \left(\sum_{i=1}^{A} h_i \mathbf{1}_{P_i}^\top \Sigma_i^{-1} \mathbf{1}_{P_i} \right) \begin{bmatrix} T \\ \mathbf{1}_T \end{bmatrix}$$

$$= \left(\sum_{i=1}^{A} \mathbf{1}_{P_i}^{\top} \Sigma_i^{-1} \mathbf{1}_{P_i} \right)^{-1} \left(\sum_{i=1}^{A} h_i \mathbf{1}_{P_i}^{\top} \Sigma_i^{-1} \mathbf{1}_{P_i} \right) \mathbf{K}^{\top} \begin{bmatrix} 0 \\ \mathbf{1}_T \end{bmatrix} = \mathbf{0}_{T-1}.$$

For the multivariate case we add a new index ν for the additional factor

$$y_{tilr}^{p,\nu} = \mu^{p,\nu} + \beta_t^{\nu} + h_i^{\nu} + u_{ti}^{1,\nu} + u_{til}^{2,\nu} + e_{tilr}^{\nu},$$

with $\nu = 1, ..., d$ and the calculations with Kronecker products run similarly to those of equation (2.9).

Using the matrix formulation, the linear mixed model is rewritten as

$$\mathbf{y}^p = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \mathbf{Z}_1\mathbf{u}^1 + \mathbf{Z}_2\mathbf{u}^2 + \mathbf{e},$$

where $\mathbf{y}^p$ is the previous data vector, $\boldsymbol{\beta} = [\mu, \beta_1, \ldots, \beta_T]^{\top}$ is the vector of fixed effects, $\mathbf{h}$ is the vector of biases, $\mathbf{u}^1$ and $\mathbf{u}^2$ are vectors containing random effects of pools and mixtures and $\mathbf{e}$ contains the residuals. The random effects are attached to the belonging measurements by $\mathbf{Z}_1$ and $\mathbf{Z}_2$ and their covariances are functions of the pool size. The corresponding weights for the covariances of $\mathbf{u}^1$ are the entries of the diagonal matrix $\mathbf{G}_1$ and equal to the inverse pool sizes. $\frac{\gamma_i - 1}{\gamma_i^2}$ are the elements of the diagonal matrix $\mathbf{G}_2$, matching to every mixture of a pool of size γ_i. Therefore the matrices $\mathbf{Z}_1$ and $\mathbf{Z}_2$ have the dimensions $n \times TA$ and $n \times T\Lambda$ respectively, with $\Lambda = \sum_{i=1}^{A} L_i$. Then, the covariance matrix has the form

$$\mathbf{V} = \mathbf{Z}_1\mathbf{G}_1\mathbf{Z}_1^{\top}\sigma_1^2 + \mathbf{Z}_2\mathbf{G}_2\mathbf{Z}_2^{\top}\sigma_2^2 + \mathbf{I}_n\sigma_t^2. \tag{2.12}$$

We obtain the estimators for the fixed and random effects from the mixed model equations

$$\begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z}_1 & \mathbf{X}^\top\mathbf{Z}_2 \\ \mathbf{Z}_1^\top\mathbf{X} & \mathbf{Z}_1^\top\mathbf{Z}_1 + \mathbf{G}_1^{-1}\lambda_1 & \mathbf{Z}_1^\top\mathbf{Z}_2 \\ \mathbf{Z}_2^\top\mathbf{X} & \mathbf{Z}_2^\top\mathbf{Z}_1 & \mathbf{Z}_2^\top\mathbf{Z}_2 + \mathbf{G}_2^{-1}\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}^1 \\ \hat{\mathbf{u}}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top\mathbf{Y} \\ \mathbf{Z}_1^\top\mathbf{Y} \\ \mathbf{Z}_2^\top\mathbf{Y} \end{bmatrix}$$

where $\lambda_1 = \frac{\sigma_t^2}{\sigma_1^2}$ and $\lambda_2 = \frac{\sigma_t^2}{\sigma_2^2}$. Since the variance parameter $\sigma = (\sigma_1^2, \sigma_2^2, \sigma_e^2)^\top$ is unknown, we compute first the REML-estimator $\hat{\sigma}$. The generalized least squares estimator (GLSE) for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \{\mathbf{X}^\top\mathbf{V}(\hat{\sigma})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^\top\mathbf{V}(\hat{\sigma})\mathbf{Y}$. The estimator $\hat{\sigma}$ introduces extra variability to $\hat{\boldsymbol{\beta}}$, making adjustments in the covariance matrix $\hat{\boldsymbol{\Phi}}_A$ of $\hat{\boldsymbol{\beta}}$ necessary.

We test for differences in gene expression between treatments with the help of a scaled F-test with approximative denominator degrees of freedom m as suggested in Kenward and Roger (1997) $F^* = \frac{m}{m+(T-1)-1}\frac{1}{T-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{K}(\mathbf{K}^\top\hat{\boldsymbol{\Phi}}_A\mathbf{K})^{-1}\mathbf{K}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Various designs can be employed depending on the question of interest for the researchers. Therefore we introduce a modified e-optimality criterion based on Landgrebe *et al.* (2004). The matrix $\mathbf{K}^\top$ stands for T-1 experimental questions and, since the covariance matrix of our generalized least squares estimator is $\hat{\boldsymbol{\Phi}}_A$, we have $\mathrm{Cov}(\mathbf{K}^\top\hat{\boldsymbol{\beta}}) = \mathbf{K}^\top\hat{\boldsymbol{\Phi}}_A\mathbf{K}$. Following their minimax approach, we set the e-efficiency to $\mathrm{tr}\left(\mathbf{K}^\top\mathbf{K}\right)[\lambda_{\max}(\mathbf{K}^\top\hat{\boldsymbol{\Phi}}_A\mathbf{K})]^{-1}$ and are able to calculate the efficiency of our design. In general, one needs estimates of variance components to choose the most efficient design. Optimal designs were considered in Passos *et al.* (2009), where Fisher's information matrix was used as criterion for the e-efficiency. For the practical decision regarding optimal design, expense-effectiveness is assessed with the help of a cost-function.

In the Appendix we compare the power of the new approach for the individual sample, equally sized pools and flexible pooling design. In the first three experiments each sample was only used once, so that no correlations or replications of the pools resulted. As our method includes more individuals, it yields smaller biological vari-

ances. This also reduces the variances of the contrast functions and leads, in theory, to higher precision of the corresponding tests. We verified this with simulated data and illustrated the power of our F*-test for DEGs in a heteroscedastic linear model.

2.2.4 Repeated use of individuals in pool building

In certain cases individuals may contribute RNA to more than a single pool. An example from the literature is the data set analyzed in Kendziorski *et al.* (2005), where 24 individuals were used to build pools of the sizes 2, 3 and 12. For a joint analysis of all pools (including individuals as pools of size one) correlations between pools are reflected in the matrix $\mathbf{G}_1$. Further, if RNA cannot be obtained from one individual - e.g. because of lost antenna of insects - it's RNA can sometimes be replaced by that from an individual from another pool. In that case the conditions for unbiasedness maintained, though at the cost of correlations between pools sharing RNA from joint individuals. For correct handling of resulting correlations we alter our assumption that each individual appears in one pool only, and we introduce the vectors

$$\mathbf{b}_{tj}(q) = \begin{cases} 1, \text{ if individual } q \in \text{pool } (t,j) \\ 0, \text{ otherwise} \end{cases}, \text{ where } q = 1, \ldots, N.$$

We assume that the S pools do not necessarily have different sizes $\gamma_1, \ldots, \gamma_S$ in each treatment and we do not consider any repeated mixtures or technical replications. The true gene expression level of a gene in a pool (t,j) is now written as

$$m^p_{tj} = \sum_{q=1}^{N} \frac{z_{tjq}\mathbf{b}_{tj}(q)}{\sum_{s=1}^{N} z_{tjs}\mathbf{b}_{tj}(s)} m_{tq} = \sum_{q=1}^{N} w_{tjq} m_{tq},$$

where z_{tjq} are independent and normally distributed $\mathrm{N}(1, \sigma_z^2)$ random variables. The weights w_{tjq} have expectation $\mathrm{E}(w_{tjq}) = \mathbf{b}_{tj}(q)\frac{1}{\gamma_j}$ and variance $\mathrm{Var}(w_{tjq}) \approx$

$\mathbf{b}_{tj}(q)\frac{\gamma_j-1}{\gamma_j^3}$. By using the two-dimensional Delta method we approximate

$$\mathrm{Cov}\left(\log(m^p_{tj_1}),\log(m^p_{tj_2})\right) = \frac{1}{\mathrm{E}(m^p_{tj_1})\mathrm{E}(m^p_{tj_2})}\mathrm{Cov}(m^p_{tj_1},m^p_{tj_2}),$$

where $j_1 \neq j_2,\ j_1, j_2 \in \{1,\ldots,S\}$. We now compute

$$\begin{aligned}\mathrm{Cov}\left(m^p_{tj_1},m^p_{tj_2}\right) &= \sum_{q=1}^{N}\frac{1}{\gamma_1\gamma_2}\mathbf{b}_{tj_1}(q)\mathbf{b}_{tj_2}(q)\mathrm{Var}(m_{tq}) \\ &\approx \frac{\mathbf{b}_{tj_1}*\mathbf{b}_{tj_2}}{\gamma_1\gamma_2}\left(e^{2\mu_t+2\sigma_b^2}-e^{2\mu_t+\sigma_b^2}\right),\end{aligned}$$

where $\mathbf{b}_{tj_1}(q)*\mathbf{b}_{tj_2}(q)=\sum_{q=1}^{N} b_{tj_1}(q)b_{tj_2}(q)$ is the number of individuals belonging to both pools. Finally it holds

$$\mathrm{Cov}\left(\log(m^p_{tj_1}),\log(m^p_{tj_2})\right) \approx \frac{\mathbf{b}_{tj_1}*\mathbf{b}_{tj_2}}{\gamma_1\gamma_2}\sigma_1^2.$$

The covariances between the logarithm of the true gene expression levels of the pools are proportional to the number of individuals used in both pools. These correlations between the pools are introduced in the matrix $\mathbf{G}_1$ and change its diagonal structure as we illustrate with the example proposed in the Appendix (see figure 3). This leads to a much more complex structure of the covariance matrix $\mathbf{V}$, hence calculations of the inverse of $\mathbf{V}$ and the GLSE could be computationally expensive. On the other hand, ignoring these correlations may cause errors in the estimation.

2.2.5 Two-color arrays

In this section we apply our model to the case of two-color arrays, where mRNA of two samples is compared by dyeing them in green and red sepia. This microarray methodology utilizes incomplete block designs and allows to discard the array and dye effects, modeling only differences between samples. Dye swap strategies lead to correlations as these designs include repeated measurements for the samples. If an

experiment involves only two groups, direct comparisons provide more information regarding a specific contrast (Rosa *et al.* (2005)). With increasing complexity of the experiment, a setup which eliminates the pooling biases is more difficult to realize. The basic idea is to apply the above system of one-color arrays. As the so called M-value is the log-ratio of two intensities corresponding to the samples of an array, we build a vector of differences by means of the matrix $\mathbf{D}$. $\mathbf{D}$ has a row for each two-color array with the entries 1 and -1 for the two selected pools and zero otherwise. The simple example of putting two consecutive pools on an array is shown in the Appendix. We obtain the model $\mathbf{M} := \mathbf{Dy} = \mathbf{DX\beta} + \mathbf{Dh} + \mathbf{e}$.

Hence, the structure of the covariance matrix $\mathbf{V}$ changes to $\mathbf{DVD}^\top$. We have $\mathrm{E}(\mathbf{M}) = \mathbf{DX\beta} + \mathbf{Dh}$. If $\mathbf{K}^\top\hat{\boldsymbol{\beta}} = \mathbf{K}^\top\left[(\mathbf{DX})^\top\mathbf{V}^{-1}\mathbf{DX}\right]^-(\mathbf{DX})^\top\mathbf{V}^{-1}\mathbf{y}$ is estimable, it holds

$$\begin{aligned}\mathrm{E}\left[\mathbf{K}^\top\hat{\boldsymbol{\beta}}\right] &= \mathbf{K}^\top\left[(\mathbf{DX})^\top\mathbf{DX}\right]^-(\mathbf{DX})^\top\mathbf{D}(\mathbf{X\beta}+\mathbf{h})\\ &= \mathbf{K}^\top\boldsymbol{\beta} + \mathbf{K}^\top\left[(\mathbf{DX})^\top\mathbf{DX}\right]^-(\mathbf{DX})^\top\mathbf{Dh}.\end{aligned}$$

It follows that $\mathbf{Dh} = \mathbf{0}$ is a sufficient condition for unbiasedness.

We notice that the compensation of biases in $\mathbf{K}^\top\hat{\boldsymbol{\beta}}$ happens simultaneously with taking differences if pool size does not vary inside the blocks. That means, we recommend hybridizing only two pools of the same size together on an array.

If dye swap and dye balance strategies are used to cancel out the dye bias (e.g. Knapen *et al.* (2009)), our mixed models handle the resulting correlations in the measurements in an appropriate way. In the described honey bee experiment, dye balance is applied, which means that pools of rare cases are labeled an equal number of times with both the red and green dye. In odd groups we hybridize one pool from the rare cases group (with swapped color) a second time, but together with a new control pool, which reduces correlations and improves the power compared to a standard dye swap.

2.3 Discussion

2.3.1 Designs with flexible pooling

First, we investigate the bias in pooled experiments in the scenario of one-color arrays. Flexible pool sizes are useful in experiments which compare subjects with a not predetermined size of the treatment groups. Additional subjects can be included by increasing the size of pools in the study without increasing costs for arrays. Our statistical analysis begins with approximative bias and variance of a pool which were computed in Zhang *et al.* (2007). The bias depends on the size of the pool, but does not tend to zero with increasing pool size γ. To enforce unbiasedness we look at the bias as a constant vector in the linear model equation, which contains this distortion for each observation. We showed, that a suitable contrast function applied to the estimated parameters can eliminate the biases after linear mapping.

Using a decomposition of the approximative variance of the pools on the log-transformed scale, we are able to deal with the occurring variance heterogeneity. Thus, we get the advantage that resulting biological variance of the pools is expected to shrink and the precision of tests should be increased. The hypothesis of no difference between gene expression levels across treatments is tested by the F^*-statistic. The variance components are unknown and can be estimated with the methods shown in Section 2.3. This increases the uncertainty, diminishing the expected gain in the accuracy of the hypotheses test for DEGs.

We investigated the publicly available data set analyzed in Kendziorski *et al.* (2005). Our method allows a joint analysis of all groups of observations, irrespective of their size. The analysis of the complete set of arrays yields a DEG ranking list, which coincides to 80% with their reference list generated from the 24 single samples. Since they used a design which included for both treatments an identical structure in the pool sizes, it fulfills the unbiasedness condition. In fact, with our criterion we can theoretically confirm their statement that '... similar amount of distortions are often observed in both control and treatment conditions ... and the bias canceled out for

most part when testing.'

Under other circumstances as in the described honeybee example it may happen, that the balance of pool sizes between treatments cannot be maintained (e.g. due to loss of samples). Then the bias has to be explicitly modeled as a cross-classified nuisance effect with one level per used pool size. The way in which the covariance matrix of all observations is modeled remains as described here and the treatment effect can then still be tested in an unbiased manner. This is, of course, at the expense of denominator degrees of freedom and power, as the model contains more fixed effects.

2.3.2 The effect of replication

We now look at the consequences of modeling mixtures in experiments with repeatedly measured pools. Suppose that all pool sizes in the experiment are equal, as recommended in the literature so far, and that for each measurement of gene expression in the laboratory we get a new mixture and do not replicate. Then it is clear from the properties of mixed models that we cannot distinguish between σ_2^2 and σ_t^2, since the structure of the random effects for mixtures and measurements will be $\mathbf{Z}_2\mathbf{G}_2\mathbf{Z}_2^\top\sigma_2^2 = \frac{\gamma-1}{\gamma^2}\mathbf{I}_n\sigma_2^2$. That simplifies the analysis of the data. Otherwise, if the pools are replicated and only mixed once, we get correlations between observations, and a distinguishable second random effect, i.e. $\mathbf{Z}_2\mathbf{G}_2\mathbf{Z}_2^\top\sigma_2^2$ is not diagonal. Then one can estimate σ_2^2 and the covariances should be regarded for the data analysis. From this viewpoint the methodology for designs with equal pool sizes has to be applied by using a new mixture for each measurement, because only in that case we get one variance component and the analysis can be done in straightforward manner. That means that reasonable analysis of the data in the general case (with replications) requires knowledge of the structure of $\mathbf{Z}_2$, making an exact documentation of each mixture of a pool in the lab necessary.

2.4 Conclusion

Due to practical challenges we developed a new, more general design for pooled gene expression experiments. The bias introduced through flexible pooling was canceled by the symmetrical structure of the pools between treatments. Hence, testing for differences between gene expression levels across treatments proves to be unbiased. To identify differentially expressed genes we choose the contrast in equation (2.8) for the linear hypothesis. Also, the provided condition allows researchers to check if a design is valid for any contrast used. Since different pool sizes cause heteroscedasticity, we propose a new statistical model comprising fixed effects (intercept and treatment effect), two random effects modeling the impact of pooling and repeated mixtures, and a technical error due to replication. In order to obtain an accurate model we advise the experimenters to make an exact documentation of each mixture of a pool. The estimation of the covariance matrix $\mathbf{V}$ is made possible by means of restricted maximum likelihood if we have different pool sizes. The variation of mixtures of the pools is underestimated in the literature and requires further analyses based on appropriate data. A new F^*-test is used which accounts for additional variability due to the estimation of the variance parameter. We apply this model to the setting of one- or two-color array experiments and we base the choice of the best design on the e-optimality approach.

Balancing the pool sizes between the treatment and control group is easily achieved in the case of two levels for the fixed effect. With every further level in the experiment the unbiasedness is still possible but takes more effort.

The presented model and methods induce substantial improvements for the design of pooled microarray experiments and allow a sound statistical analysis. The new design strategy is particularly applicable to comparisons of rare properties in large families of animals or plants for enhancing the efficiency and controlling cost.

2.5 Appendix

Approximation of biological variance of pools

We compute here the biological variance of a pool of size γ

$$\begin{aligned}\sigma_b^{2,p} \sim \frac{\mathrm{Var}\left(m_{tj}^p\right)}{\left[\mathrm{E}\left(m_{tj}^p\right)\right]^2} &= \frac{\frac{1}{\gamma}\left(e^{2\mu_t+2\sigma_b^2}-e^{2\mu_t+\sigma_b^2}\right)\left(1+\gamma^2\sigma_w^2\right)}{\left(e^{\mu_t+\sigma_b^2/2}\right)^2} \\ &= \frac{1}{\gamma}\left(e^{\sigma_b^2}-1\right)\left(1+\gamma^2\sigma_w^2\right) \\ &\approx \frac{1}{\gamma}\left(e^{\sigma_b^2}-1\right)\left(1+\frac{\gamma-1}{\gamma}\sigma_z^2\right).\end{aligned}$$

This was incorrectly displayed in Zhang *et al.* (2007) in Section 2.2 in equations (12) and (13). As a consequence the non-centrality parameter δ^2 in equation (16) is too large. That shifted the F-tests, and the power was overestimated. This error may cause the lack of response of the calculated power to pooling technical variance shown in Figure 4 of Section 3.4 in that paper. In the top of Figure 2.1 we evaluate the power in the setting of Zhang *et al.* (2007) for simplicity with only two treatments. As usual, we develop the power curves as the probability to detect a certain mean class difference ($\boldsymbol{\Delta}$) between the two treatments. On the alternative of the hypothesis $\mathbf{K}^\top\boldsymbol{\beta} = \boldsymbol{\Delta} = 0$, first the non-centrality parameter (δ^2) is calculated:

$$\delta^2 = \boldsymbol{\Delta}\left(\mathbf{K}^\top\left(\mathbf{X}^\top(\mathbf{V})^{-1}\mathbf{X}\right)^{-}\mathbf{K}\right)\boldsymbol{\Delta},$$

for both versions of the covariance matrix $\mathbf{V}$. For $\mathbf{X}$ we select $\begin{bmatrix}1&1&0\\1&0&1\\&\vdots&\end{bmatrix}$ and take the suitable contrast function $\mathbf{K}^\top = [0\,1\,\text{-}1]$. Then the power of the F-test to detect DEGs can be determined according to:

$$1-q = 1-F_{\delta^2,1,2*(S-1)}(f_{0.95,1,2*(S-1)})$$

Checking numerically the relative change from the corrected power we find distortions of up to 4%. For the theoretical power calculations in the bottom of Figure 2.1

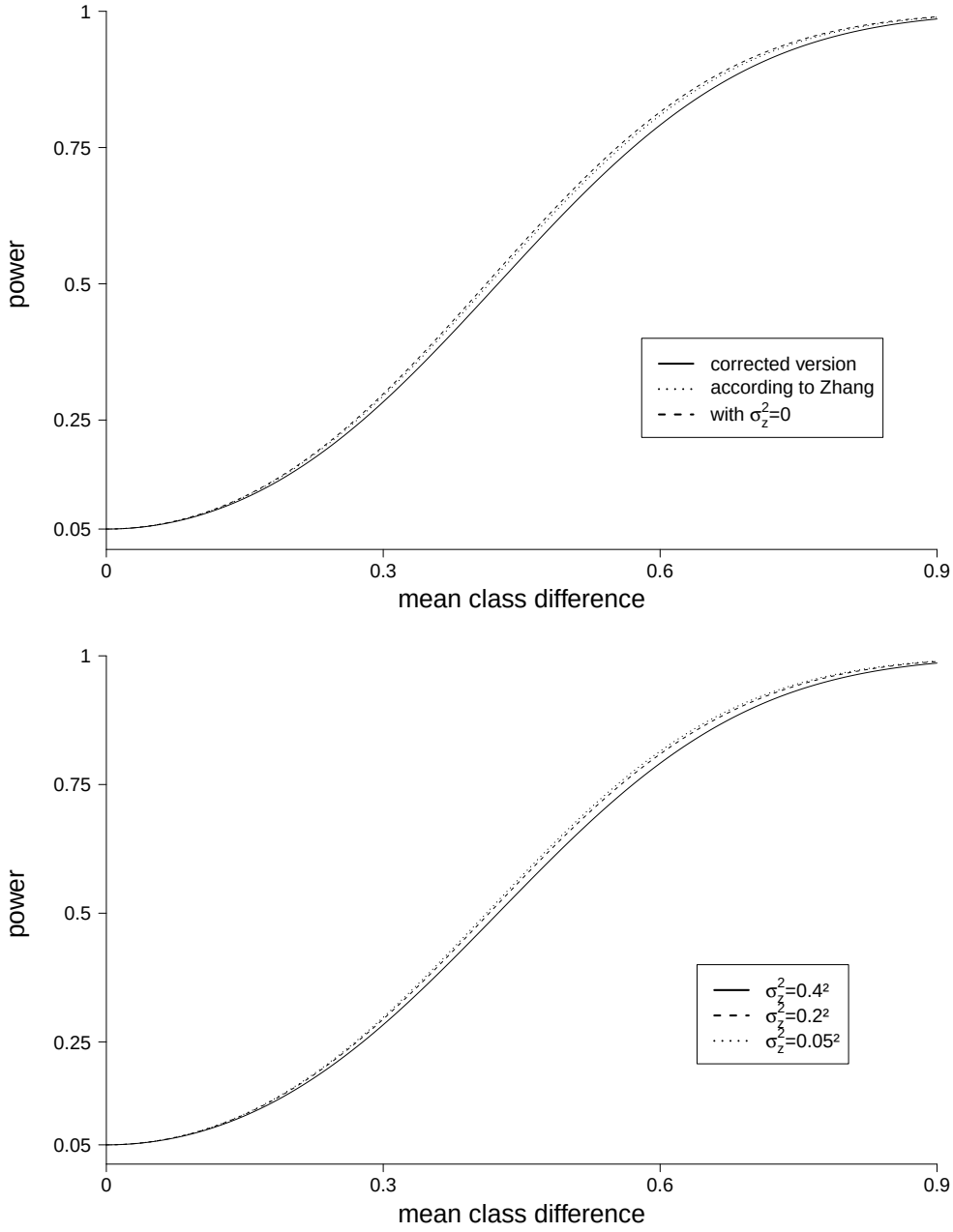


Figure 2.1: Power for the detection of a difference between two treatments in an experiment with 100 individuals per treatment, measured in 25 pools of size 4. Top: power curve for $\sigma_z^2 = 0.4^2$ when the correct formula is used (solid) and with the formula of Zhang (dotted). The latter almost coincides with the curve for $\sigma_z^2 = 0$ (dashed). Bottom: power curves for $\sigma_b^2 = 0.75$, $\sigma_t^2 = 0.25$ and $\sigma_z^2 \in \{0.4^2, 0.2^2, 0.05^2\}$.

different levels of the pooling technical variance are used. The slope of the power curves is weaker with increasing σ_z^2, as expected.

Variance structure for a two-color experiment

Generally, two-color microarray experiments are illustrated by a set of arrows (e.g. Young and Speed (2002)). Their tail and head denote the green and red labeling assignments. We look at an experiment with four stochastically independent pools, four two-color arrays and two pool sizes 2 and 4 (see Figure 2.2). We build the

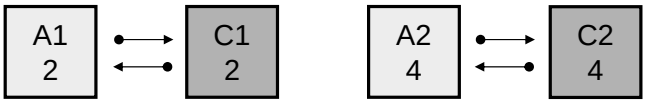


Figure 2.2: Scheme for a two-color experiment with four independent pools, two different pool sizes and four arrays

matrix $\mathbf{G}_1$ with the inverse pool sizes on the diagonal to estimate the biological variance,

$$\mathbf{G}_1 = \begin{bmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}. \quad \text{The matrix } \mathbf{Z}_1 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

is the design matrix for the random effect of the pools. If we have eight mixtures, then $\mathbf{G}_2$ contains eight entries $\frac{\gamma_i - 1}{\gamma_i^2}$ on its diagonal. Then $\mathbf{Z}_2 = \mathbf{DI}_8$ holds, with

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

From equation (2.12) we get:

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 + \frac{1}{2}\sigma_2^2 + \sigma_t^2 & -\sigma_1^2 & 0 & 0 \\ -\sigma_1^2 & \sigma_1^2 + \frac{1}{2}\sigma_2^2 + \sigma_t^2 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\sigma_1^2 + \frac{3}{8}\sigma_2^2 + \sigma_t^2 & -\frac{1}{2}\sigma_1^2 \\ 0 & 0 & -\frac{1}{2}\sigma_1^2 & \frac{1}{2}\sigma_1^2 + \frac{3}{8}\sigma_2^2 + \sigma_t^2 \end{bmatrix}.$$

Unbiasedness for two-factorial experiment

We denote the factor of interest with A and the disturbance factor as D. In the minimal case we have a crossed factorization with two levels for each factor. We draw one box for each pool including factors and pool sizes. The pool sizes of the

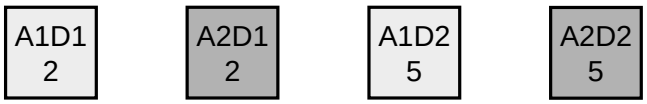


Figure 2.3: Scheme for an experiment with four independent pools, two different pool sizes and two factors

levels of the first factor are equal on each level of the second factor (see Figure 2.3). Then we use $\mathbf{K}^\top = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \end{bmatrix}$.

The covariance matrix is $\mathbf{V} = \begin{bmatrix} v_1 & 0 & 0 & 0 \\ 0 & v_1 & 0 & 0 \\ 0 & 0 & v_2 & 0 \\ 0 & 0 & 0 & v_2 \end{bmatrix}$. Since it holds

$$\left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^- \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{h} = \begin{bmatrix} \frac{2}{v_1}+\frac{2}{v_2} & \frac{1}{v_1}+\frac{1}{v_2} & \frac{1}{v_1}+\frac{1}{v_2} & \frac{2}{v_1} & \frac{2}{v_2} \\ \frac{1}{v_1}+\frac{1}{v_2} & \frac{1}{v_1}+\frac{1}{v_2} & 0 & \frac{1}{v_1} & \frac{1}{v_2} \\ \frac{1}{v_1}+\frac{1}{v_2} & 0 & \frac{1}{v_1}+\frac{1}{v_2} & \frac{1}{v_1} & \frac{1}{v_2} \\ \frac{2}{v_1} & \frac{1}{v_1} & \frac{1}{v_1} & \frac{2}{v_1} & 0 \\ \frac{2}{v_2} & \frac{1}{v_2} & \frac{1}{v_2} & 0 & \frac{2}{v_2} \end{bmatrix}^- \begin{bmatrix} \frac{2h_1}{v_1}+\frac{2h_1}{v_2} \\ \frac{h_1}{v_1}+\frac{h_1}{v_2} \\ \frac{h_1}{v_1}+\frac{h_1}{v_2} \\ 2\frac{h_1}{v_1} \\ 2\frac{h_2}{v_2} \end{bmatrix},$$

in the end we get $\mathbf{K}^\top \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^- \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{h} = \mathbf{0}$.

Summary of the results for simulated pooling experiments

We simulated data from two treatment groups. The number of pools was fixed to 24. The pool sizes were chosen randomly from a shifted Poisson distribution ($\lambda = 3$) with expectation $3 + 2$. For the individual sample design we chose the pool size 1. For the second design the pool size was determined as the minimum size of all pools from the treatment group. In the third design we applied the new principle for flexible pooling designs: balanced sizes of pools between treatment and control

group. In the fourth design we reduced the pools to twelve and added one technical replication for each. Thereby we evaluated correlations with a comparable number of used individuals to the second design. Here the choice of the technical error had a strong influence on the results. We assumed $\mathbf{V}$ to be known with the variance components $\sigma = (\sigma_b^2 = 0.12, \sigma_z^2 = 0.1^2, \sigma_t^2 = 0.2^2)$ and used them to generate the biological and technical errors. For the designs 1,2 and 3 we assumed to have independent pools and no replications, so that $\mathbf{V}$ is diagonal. The mean class difference was estimated with the GLSE. Then we applied an F-test to decide whether a gene

pool design	estimation of contrast	est. var contrast	theor. power	simul. power	avg. no. used individuals
single	0.400	0.027	0.6486	0.649	24
equal	0.401	0.015	0.8700	0.876	63
flexible 1	0.401	0.011	0.9488	0.950	120
flexible 2	0.399	0.016	0.8603	0.859	60

Table 2.1: Results of simulation study for comparison of designs with none-pooled samples, equally sized pools, flexible pool sizes without and with replication

is differentially expressed. We got the simulated power for each design as the frequency of the rejected tests in 10000 runs. Table (1) shows averaged estimations of a simulated mean class difference ($\mathbf{\Delta} = 0.4$), theoretical variance of the contrast, simulated and empirical power for DEG detection at the type I error level $\alpha = 0.05$ and the average number of used individuals. The gain in accuracy is visible and most valuable if the cost of the individuals are small relative to the arrays.

Construction of $\mathbf{G_1}$ for an experiment with individuals occur in several pools

In an example (see Figure 2.4) we assume 11 individuals to build four pools, where the individuals 1-4 form the first, individuals 3-6 the second, 7-9 the third and 9-11 the fourth pool. The matrix $\mathbf{G}_1^*$ includes the pool sizes on the diagonal $g_{11} = 4, g_{22} = 4, g_{33} = 3, g_{44} = 3$ and also the number of shared individuals at $g_{12} = g_{21} = 2$ and $g_{34} = g_{43} = 1$. Then we calculate

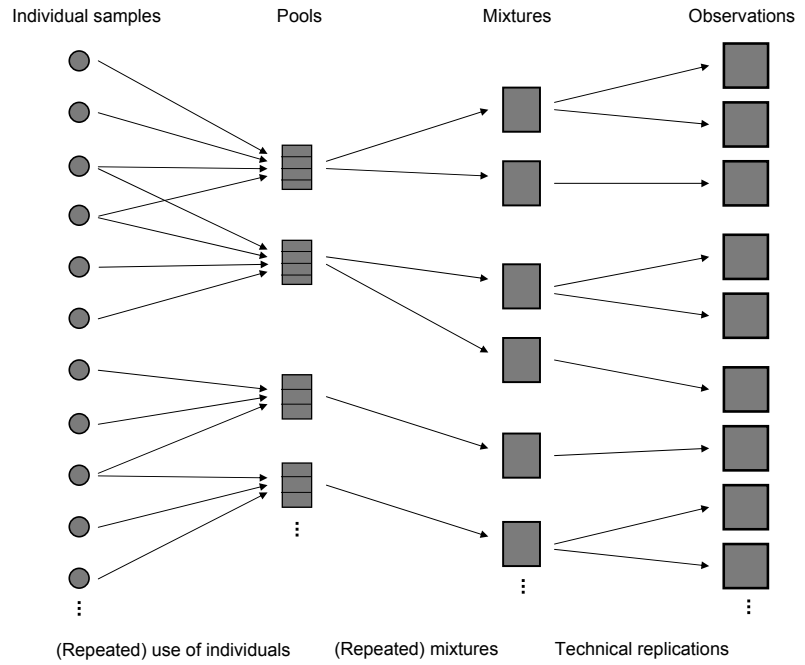


Figure 2.4: Scheme for possibilities of replication

$$\mathbf{G}_1 = \mathbf{G}_1^0\mathbf{G}_1^*\mathbf{G}_1^0 = \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 4 & 2 & 0 & 0 \\ 2 & 4 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{8} & 0 & 0 \\ \frac{1}{8} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{9} \\ 0 & 0 & \frac{1}{9} & \frac{1}{3} \end{bmatrix},$$

where $\mathbf{G}_1^0$ is the diagonal matrix with the inverse pool sizes.

In general, we take $\mathbf{B}$ as the matrix built by the vectors $\mathbf{b}_{tj}$ introduced in Section 2.4 having $\mathbf{B} = [\mathbf{b}_{11}, \ldots, \mathbf{b}_{TS}]$. Then $\mathbf{G}_1$ is built by $\mathbf{G}_1 = \mathbf{G}_1^0\mathbf{B}^\top\mathbf{B}\mathbf{G}_1^0$.

3 On the relevance of technical variation due to building pools in microarray experiments

3.1 Introduction

In gene expression profiling pooling is a method to reduce hybridization costs and compensate for insufficient amounts of mRNA. In the subsequent statistical analyses of gene expression data, where a log-transformation during preprocessing is standard, it is important to consider how the expectation and variance of the gene expression of pools relate to individual samples. The impact of pooling on the identification of differential gene expression has been studied in Kendziorski *et al.* (2005), separately for different pool sizes. It has been shown that biological averaging occurs for most of the transcripts and differential expression inferences are comparable for individuals and pools. In Zhang *et al.* (2007) approximations for the expectation and variance of pooled samples were derived. Furthermore, it was shown that biases as well as heteroscedasticity are introduced by variable pool sizes. Experiments with unequal pool sizes therefore were recommended to be avoided. As demonstrated in Rudolf *et al.* (2013), however, a wide class of experiments, in which pool size can be handled as a nuisance effect and is cross-classified with treatment, allows for tests of unbiased contrasts. In the case of a balanced cross-classification the pool size effect

must not explicitly appear in the model at all, though hypotheses on treatments remain unbiased, as shown in Rudolf *et al.* (2013). In any case variable pool sizes have an effect on the covariance of observations. This can be taken into account by considering how many individuals are allocated to each pool and by introducing a random effect for blending along with a corresponding variance component. The latter can be interpreted as a second kind of technical variability induced by inaccuracies in blending slightly unequally-sized aliquots of mRNA from several individuals into common pools. Though this subject has been treated theoretically as described, investigation into the practical importance of this second kind of technical variability is lacking.

Consequently a study was performed, in which gene expression data from experiments with four different species were analyzed to investigate the relevance of the aforementioned new kind of technical error in terms of size and significance of the corresponding variance component. Furthermore, we investigated potential consequences on the number of transcripts identified as differentially expressed between different treatments when analyses neglect this kind of error.

3.2 Material and Methods

This section offers a short recap of the underlying statistical models. The four experimental data sets are then introduced. In all of them - whether from single-

Characteristics	Mouse	Rat	Bee	Human
Individuals	60	24	14	55
Pools	12	22	12	16
Pool size	5	2,3,12	2,4	3
Observations	44	56	22	30
1-/2- color-array	1	1	2	2

Table 3.1: Overview of properties of the experimental data sets.

or two-color arrays - there are more observations than pools (see Table 3.1), which allows for the estimation of all desired variance components. Data simulations are

also described and have been included as a useful aid for the interpretation of the experimental data results. Finally, the statistical methods applied for parameter estimation and statistical testing are described.

3.2.1 Random effects in gene expression experiments with variable pool sizes

When aliquots of mRNA from different individuals are blended into common pools, the inaccuracies of this procedure may induce a special kind of technical error. Respective random effects, together with a corresponding variance component, were proposed (Rudolf *et al.*, 2013) as a means of modeling the variability of pooled observations in gene expression experiments with variable pool sizes (i.e. differing numbers of individuals per pool). Thus, for background-corrected and normalized log-intensities $\mathbf{y}$ (length of vector $\mathbf{y}$ equals the number of arrays) of a certain transcript, the model in matrix notation is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}, \tag{3.1}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices of the fixed ($\boldsymbol{\beta} = (\mu, \beta_t)^\top$) and random ($\mathbf{u}_1, \mathbf{u}_2$) effects. The distribution of $\mathbf{u}_j$ is assumed to be $\mathbf{u}_j \sim N(\mathbf{0}, \mathbf{G}_j\sigma_j^2),\ j = 1, 2$ with covariance matrices $\mathbf{G}_j\sigma_j^2$ (σ_j^2 are the variance components) and the residuals are $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. Random effects of single individuals are assumed to be independently identically distributed with a biological variance σ_1^2, while observations from a number of γ_i pooled individuals have a biological variance $\frac{\sigma_1^2}{\gamma_i}$. The vector $\mathbf{u}_1$ may comprise biological effects of single individuals as well as average biological effects of groups of individuals constituting common pools, according to the experimental design.

The random effect of blending (i.e. for the technical procedure of building a pool) only applies to observations from pools and not to observations from single individu-

als. Therefore, $\mathbf{u}_2$ consists of one effect per mixture, which had been prepared in the lab. The associated variance component is σ_2^2. So, the variance of the observations becomes:

$$\mathbf{V}(\mathbf{y}) = \mathbf{Z}_1\mathbf{G}_1\mathbf{Z}_1^\top\sigma_1^2 + \mathbf{Z}_2\mathbf{G}_2\mathbf{Z}_2^\top\sigma_2^2 + \mathbf{I}_n\sigma_e^2. \tag{3.2}$$

The model of this variance structure is based on the closed form approximation of the variance of pools on the scale of log-intensities, proposed in Zhang *et al.* (2007)

$$v_i \approx \left(e^{\sigma_b^2} - 1\right)\frac{1}{\gamma_i} + \left(e^{\sigma_b^2} - 1\right)\sigma_z^2\frac{\gamma_i - 1}{\gamma_i^2}, \tag{3.3}$$

where σ_z^2 is the pooling technical variance and σ_b^2 is the biological variance of individuals. The substitutions $\sigma_1^2 := e^{\sigma_b^2} - 1$ and $\sigma_2^2 := (e^{\sigma_b^2} - 1)\sigma_z^2$ led to our assumed variance structure (3.2).

In the following, the relevance of accounting for the blending error variance component σ_2^2 is investigated in four experimental data sets by comparing the described full model (m2) described above with a reduced one (m1) that lacks this particular variance component. The methodology was checked by a simulation beforehand.

3.2.2 Experimental data

Mouse data

Mouse data consisted of observations from 44 one-color microarrays. RNA for this experiment was extracted from the ovaries of 60 female mice, 30 of which came from a long-term selection line with an extraordinary litter size. All others came from a control line. Pooled samples were built by blending RNA from five mice per sample. Each mouse was only represented in a single pool. For the sake of technical replication, all 12 pooled samples were measured twice by preparing two microarrays per sample. Additionally, animals from two pools per line (10 animals per line) were measured individually. These individual measurements were not included in

the previously published analysis of this data (Vanselow *et al.*, 2008), where more details of the experiment can be found.

28 (14 per line) different biological effects were defined per transcript. The dimensions of the design matrix $\mathbf{Z}_1^m$ are therefore 44×28. In detail, random biological effects were assigned to all individually measured mice (individuals 1 to 10 within each line) and corresponding entries in $\mathbf{Z}_1^m$ equal 1. The biological effects of the same ten individuals (in two groups of five) were assigned to the observations from the first and second pooled samples in each line (two observations per pool due to technical replication). In this case, non-zero entries in $\mathbf{Z}_1^m$ are 1/5. However, pooled samples numbered 3 to 6 within each line each had a biological effect of their own, modeling the average effect of the five respective members of each pool. Note that for the pools 3 to 6 the corresponding non-zero entries in $\mathbf{Z}_1^m$ are 1. The 28×28 covariance matrix $\mathbf{G}_1^m$ scales the random biological effects and has non-zero entries only on the diagonal, each of them equals the inverse pool size $1/\gamma_i$. The 22 observations from the first line are represented in the upper part of the design matrix $\mathbf{Z}_1^m$:

$$
\mathbf{Z}_1^m = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & & & & \vdots & & & & & & \vdots & & & & & & & \vdots & &
\end{bmatrix}.
$$

The technical variability due to blending individual samples only comes into play when observing pooled samples, not for measurements of individuals. Since blending was done only once per pool, there are 12 different effects due to imperfect blending. Therefore, the 44×12 design matrix $\mathbf{Z}_2^m$ (see section 7.1.2) contains zero rows for observations from single animals. The corresponding 12×12 covariance matrix $\mathbf{G}_2^m$ is diagonal with entries $\frac{\gamma-1}{\gamma^2} = \frac{4}{25}$, according to equation (3.3).

Rat data

This data set is publicly available at the Gene Expression Omnibus database website (accession no. GSE2331) and contains one-color array data. Rats of the treatment group were treated with Retinoic acid. For the details of data generation and preprocessing, please see the original paper from Kendziorski *et al.* (2005). Rats from the groups A (control) and B (treatment) were measured individually and in pools of various sizes. Each of the twelve rats from both groups was used 4 times, for an individual measurement and in pools of 2, 3, and 12. For the sample composition we again defined the random effects from the smallest disjunct elements. Therefore, with the help of the matrices $\mathbf{G}_1^r$ and $\mathbf{Z}_1^r$, convex linear combinations were built from the 24 individuals. Here, $\mathbf{G}_1^r$ is the 24×24 unity matrix and $\mathbf{Z}_1^r$ contains a row for each measurement with entries according to reciprocal pool sizes. Per group, there are 28 measurements partitioned into 12 individual samples, 6 pools of 2, 4 pools of 3, and one of 12, plus 5 technical replications. Thus, the dimensions of the matrix $\mathbf{Z}_1^r$ are 56×24, detailed in section 7.1.2. In each group, there were 11 pools, and the diagonal matrix $\mathbf{G}_2^r$ has the dimensions 22×22 with entries $\{\frac{2}{9}, \frac{1}{4}, \frac{2}{9}, \frac{1}{4}, \frac{2}{9}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{2}{9}, \frac{1}{4}, \frac{11}{144}, ...\}$. The matrix $\mathbf{Z}_2^r$ was constructed analogously to $\mathbf{Z}_1^r$.

Honeybee data

This data set stems from a honeybee project dealing with differences in the pathogen resistance of so-called hygienic and non-hygienic worker bees as far as they are reflected in gene expression differences. Bees designated as 'hygienic' were observed to open brood cells and assisting the removal of diseased brood. The bees' activities were recorded on a *Varroa*-parasitized section of a brood comb. Pooling was applied in a preliminary experiment with a limited number of bees and microarrays. For 7 hygienic bees and 7 control bees, mRNA was extracted from nerve tissues of the mushroom body (MB), antennal lobe (AL) and Antennae (ANT). The number of individuals blended into a pool was either 2 or 4. Out of the 14 bees, 6 different

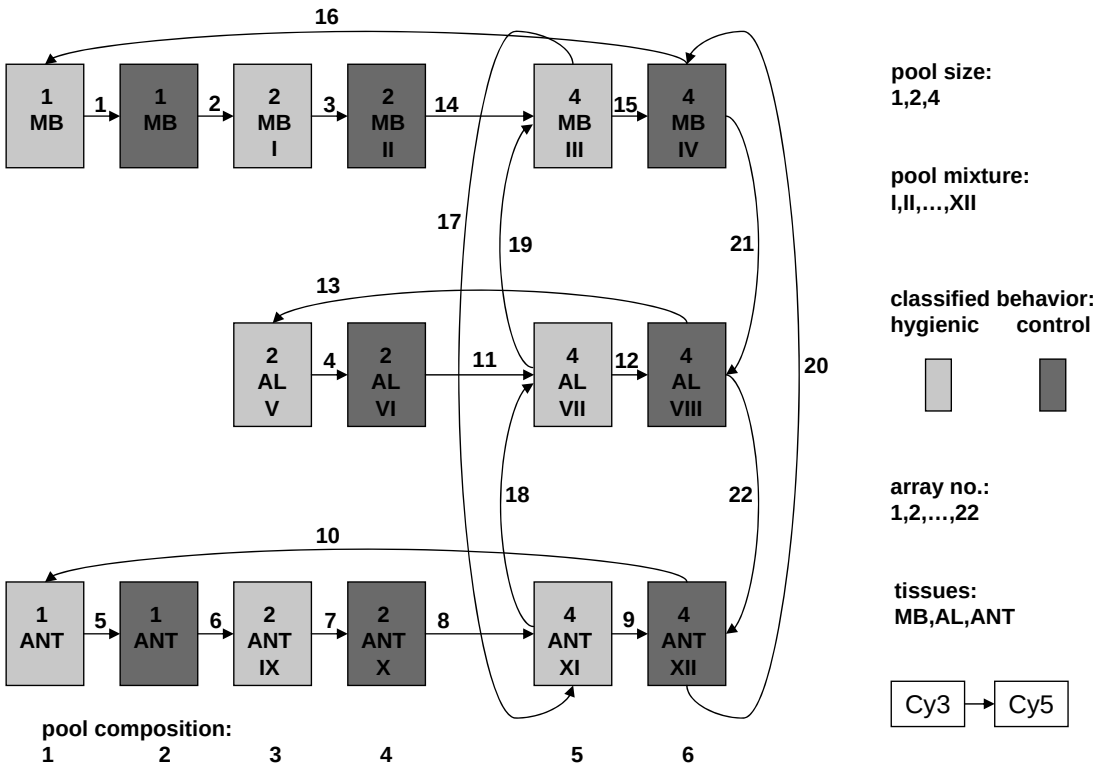


Figure 3.1: Scheme for the design of the two-color microarray experiment with honeybees. The numbered arrows (1-22) represent two-color arrays, the arrowheads (tails) indicate Cy5 (Cy3) dye. Light (dark) boxes symbolize RNA from hygienic (control) bees. Pool size (1, 2, 4) and mixture (Roman numerals) are shown in each box. Tissues are abbreviated as MB (mushroom body), AL (antennal lobe), and ANT (Antennae). Boxes in the same column share the same biological effect, indicated as pool compositions 1 to 6.

sample compositions were built and analyzed for all three tissues with two-color arrays (for the design see Figure 3.1). A few individual hybridizations were not carried out due to an insufficient amount of amplified RNA (single samples from AL). For the normalized two-color microarray data we used a model for differences $\mathbf{M}$ of log-intensities from the red (R) and green (G) channel

$$\mathbf{M} = \mu + \boldsymbol{\Delta} + b_{12} + b_{23} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}. \tag{3.4}$$

Here $\mathbf{M}$ is the vector of log-ratios ($\mathbf{M} = \log\frac{R}{G} = \log R - \log G$) for one transcript with dimension n, equal to the number of arrays. The design matrix $\mathbf{X}$ for the fixed effects links observations to the overall mean μ (which includes the dye effect, i.e.

the difference of red and green channel), the differences $\boldsymbol{\Delta}$ between the behaviors (hygienic minus control) and two differences between tissues (b_{12} for MB minus AL, b_{23} for AL minus ANT). The latter effect has been included since data from all tissues were jointly analyzed due to the limited number of arrays. The random effect $\mathbf{u}_1$ for each sample composition has a variance structure determined by $\mathbf{G}_1^h$ and $\mathbf{Z}_1^h$. The variance structure of the second random effect $\mathbf{u}_2$ for the blending of individuals is generated by $\mathbf{G}_2^h$ and $\mathbf{Z}_2^h$. Both design matrices for the random effects differ, however, from experiments with one-color arrays: each row of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ contains two non-zero elements (as opposed to a single one) in order to model the differences between effects with entries of 1 for the red and -1 for the green channel. The residual errors $\mathbf{e} \sim N(0, \sigma_e^2)$ are again assumed to be stochastically independent and include the technical errors created through the hybridization, imaging, and scanning of each array.

Human data

The human data was taken from the GC6 (Grand Challenge in Global Health no. 6 - Biomarkers of protective immunity against Tuberculosis) project. For the project data, please see Maertzdorf *et al.* (2010). One focus of this project was to identify immune system differences between people who were exposed to Tuberculosis but never became sick and those who developed severe symptoms. Therefore, as a part of this larger study the three classes TST^+, TST^- and TB were compared, where TST stands for the tuberculosis skin test (+ and - indicate positive and negative results, respectively) and TB for acute tuberculosis. Overall, the data set consists of samples from 55 humans in 16 pools of three and in 10 single samples, which were labeled on 30 two-color arrays. In the sample composition, one also sees correlations between pools in three cases, where individuals were used more than once, i.e. in different pools (see matrix $\mathbf{G}_1^g$). For each observation we modeled fixed effects for the mean (including dye effect) and treatment (3 levels) as well as random effects of sample composition and imperfect blending. Because there were two samples on

each array, the design matrix $\mathbf{Z}_1^g$ for the composition of the samples had two entries per row, as presented in section 7.1.2. Each pool was built only once, so $\mathbf{G}_2^g$ is a diagonal matrix with dimensions 16×16 and entries $\frac{2}{9}$. The random effects of imperfect blending were assigned to measurements via $\mathbf{Z}_2^g$, with two non-zero entries per measurement.

3.2.3 Simulated data

The relationship between the variance of a random effect of a pool and deviations from the homogeneous aliquots of individuals in a pool sample, given in equation (3.3), is based on a theoretically derived approximation (Zhang *et al.*, 2007). Furthermore, true proportions of aliquots are not available. Therefore, the equality of the estimated variance component σ_2^2 and the product of variances $(e^{\sigma_b^2} - 1)\sigma_z^2$ was checked by fitting the model to simulated data, in order to assay the estimations when the true state of nature is known.

By setting $\mathbf{x} \sim N(\mu_g, \mathbf{I}\sigma_b^2)$ the vector of individual gene expressions of the individuals of a pool and $\mathbf{w}$ the vector of weights (proportions of individuals in the pooled RNA of a joint sample), we calculated a value for true gene expression on the log-scale as

$$\log\left(\mathbf{w}^\top \times \exp\left(\mathbf{x}\right)\right). \tag{3.5}$$

The technical errors, distributed as $N(0, \sigma_t^2)$, were then added. Note that, due to (3.1), each observation is composed by the fixed effects $\mathbf{X}\boldsymbol{\beta} = \mu_g$, the distortion due to biological variation $\mathbf{u}_1 = \bar{\mathbf{x}} - \mu_g$ and the difference generated by imperfect blending $\mathbf{u}_2 = \log(\mathbf{w}^\top \times \exp\left(\mathbf{x}\right)) - \log\left(\overline{\exp\left(\mathbf{x}\right)}\right)$, plus the log-bias $\log\left(\overline{\exp\left(\mathbf{x}\right)}\right) - \bar{\mathbf{x}}$. For the simulation of weights the Dirichlet distribution with parameters $a_i = \frac{1}{\sigma_z^2} - \frac{1}{\gamma}$, $i = 1, \ldots, \gamma$ was used. Then, $a_0 = \sum_{i=1}^{\gamma} a_i = \gamma a_i$, and the expectation of each weight is $\frac{a_i}{a_0} = \frac{1}{\gamma}$. Therefore, the variance of the weights - theoretically $\frac{a_i(a_0 - a_i)}{a_0^2(a_0+1)}$ - is $\frac{\gamma-1}{\gamma^3}\sigma_z^2$. Using the approximation $\frac{\gamma-1}{\gamma^3}\sigma_z^2 \approx \sigma_w^2$ for the variance of weights $\mathbf{w}$ from Zhang *et al.* (2007), the Dirichlet parameters a_i can be chosen in order to obtain weights with a

given variance σ_w^2.

Various proportions of transcripts (0, 1/3, 1) were simulated as affected by imperfect blending. In order to investigate the distribution of the RLRT-statistic under the null hypothesis ($\sigma_2^2 = 0$), the pooling technical variance σ_z^2 was set to zero for all transcripts. Then, one third of the transcripts were simulated with imperfect blending, as well as data where all transcripts contained these effects.

As a test case, further simulations were tailored for a comparison of models with regard to the power to detect differential expression in the presence of imperfect pooling at all loci. Variances were set to $\sigma_t^2 = 0.17$, $\sigma_b^2 = 0.103$ and $\sigma_z^2 = 2.7$ according to the estimations from the mouse data. This was simulated with 100 repetitions. An experiment consisting of 60 individuals from two equally-sized treatment groups was simulated, in a 44 one-color microarray setting. The observations generated were both from single individuals (20) and pools of size five (24). The individual values used in the first two pools of each line were also used as single individuals. For the full details of the design, please see the description of the mouse data set above, which has an identical structure. For each of the 9000 transcripts, a mean expression level was randomly chosen from a uniform distribution over the interval $[8, 14]$. A subgroup of 3000 transcripts was randomly chosen to be differentially expressed between both treatment groups. For each of these, a mean treatment effect was sampled from a uniform distribution over the interval $[0.5, 1.5]$ with a random sign $\in \{-1, 1\}$. False positive and negative test results were then evaluated using the mean number of transcripts, averaged over all 100 repetitions.

3.2.4 Statistical analyses

Three variance components were considered: first, biological variance (σ_1^2); second, blending error variance (σ_2^2); and third, residual variance (σ_e^2). Similar models that lack the second variance component have been used previously (e.g. Yang, 2003). Transcripts were excluded from analyses if the log-expressions of both groups were

smaller than eight (corresponds to 256 at the original scale), which is frequently considered to be a threshold for meaningful gene expression. This resulted in 8554 observations for the mouse data, 6264 for rats, 13761 for bees and 12348 for the human data set. An EM-REML algorithm was used to estimate the variance components. Then the mixed model equations

$$\begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z}_1 & \mathbf{X}^\top\mathbf{Z}_2 \\ \mathbf{Z}_1^\top\mathbf{X} & \mathbf{Z}_1^\top\mathbf{Z}_1 + \mathbf{G}_1^{-1}\lambda_1 & \mathbf{Z}_1^\top\mathbf{Z}_2 \\ \mathbf{Z}_2^\top\mathbf{X} & \mathbf{Z}_2^\top\mathbf{Z}_1 & \mathbf{Z}_2^\top\mathbf{Z}_2 + \mathbf{G}_2^{-1}\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top\mathbf{Y} \\ \mathbf{Z}_1^\top\mathbf{Y} \\ \mathbf{Z}_2^\top\mathbf{Y} \end{bmatrix},$$

where $\lambda_1 = \frac{\sigma_e^2}{\sigma_1^2}$ and $\lambda_2 = \frac{\sigma_e^2}{\sigma_2^2}$, were solved for the estimates of the fixed and random effects and the REML-log-likelihood was calculated.

For each transcript, a residual likelihood ratio test (RLRT) was used to test the null hypothesis $H_0 : \sigma_2^2 = 0$, thereby assuming a half-half mixture of a χ_1^2-distribution and a point mass at zero (see e.g. Scheipl *et al.*, 2008). According to this assumed distribution of the test statistic, the distribution of p-values from all transcripts in one experiment under the null hypothesis deviates from the uniform distribution (see figure 3.2). The proportion of transcripts with a relevant blending error variance was estimated as $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Therein, the estimated proportion of true null hypotheses ($\hat{\pi}_0$) was estimated as described in Dabney *et al.* (2011). The proportion $\hat{\pi}_1$ was then compared with the proportion of transcripts simulated without blending errors. After correcting all p-values according to a false discovery rate (FDR) of 5%, the transcripts with a significant RLRT were determined. Beyond that, we evaluated the proportions of the estimated variance component σ_2^2 in relation to the total variance.

The practical relevance of the variance component for imperfect blending of samples was further investigated by comparing the number of transcripts identified as differentially expressed in different treatment levels by means of the full model (m2, equation 3.1) and the null model (m1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{e}$ without a random effect of imperfect blending. Degrees of freedom for the applied F-Tests of fixed effects in

mixed models were adjusted according to Kenward and Roger (1997). In order to account for multiple testing, an FDR of 5% was applied to the p-values of the latter F-tests.

For the mouse data set, the normalization was done with the gcrma method (Wu *et al.*, 2004). Loess- and quantile normalization (Smyth and Speed, 2003) was used for the two-color array data. The rat data set was downloaded as normalized (www.ncbi.nlm.nih.gov/geo).

The open-source statistical programming package R (R Core Team, 2012) was used to implement an EM-REML algorithm for the estimation of all three variance components. The formulas for the expectation and maximization steps can be obtained from e.g. Mrode and Thompson (2005). Convergence of the EM algorithm was assumed when the condition

$$\sqrt{\frac{\left(\mathbf{B}_{n-1}-\mathbf{B}_{n}\right)^{\top}\left(\mathbf{B}_{n-1}-\mathbf{B}_{n}\right)}{\mathbf{B}_{n}^{\top}\mathbf{B}_{n}}}<\epsilon, \tag{3.6}$$

was fulfilled Schaeffer (1986), where $\epsilon = 10^{-8}$ and $\mathbf{B}_n = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_2^2 & \hat{\sigma}_e^2 \end{bmatrix}^{\top}$ is the vector of estimates of the variance components in the *n-th* iteration. False discovery rates were computed with the help of the R-package qvalue Storey and Tibshirani (2003). In the case of p-values from RLRT test statistics, the 'bootstrap' option was used to estimate π_0, as suggested by Storey (2002).

3.3 Results and Discussion

3.3.1 Simulated data sets

First, the results of the RLRT for the blending error variance component are shown for the case of the validity of the null hypothesis ($\sigma_2^2 = 0$). Here, a uniform distribution of p-values can be observed on the interval $[0, 0.5)$ as expected (see Figure 3.2, topright). The Distributions of log-estimates of σ_2^2 (Figure 3.2, left panels from top

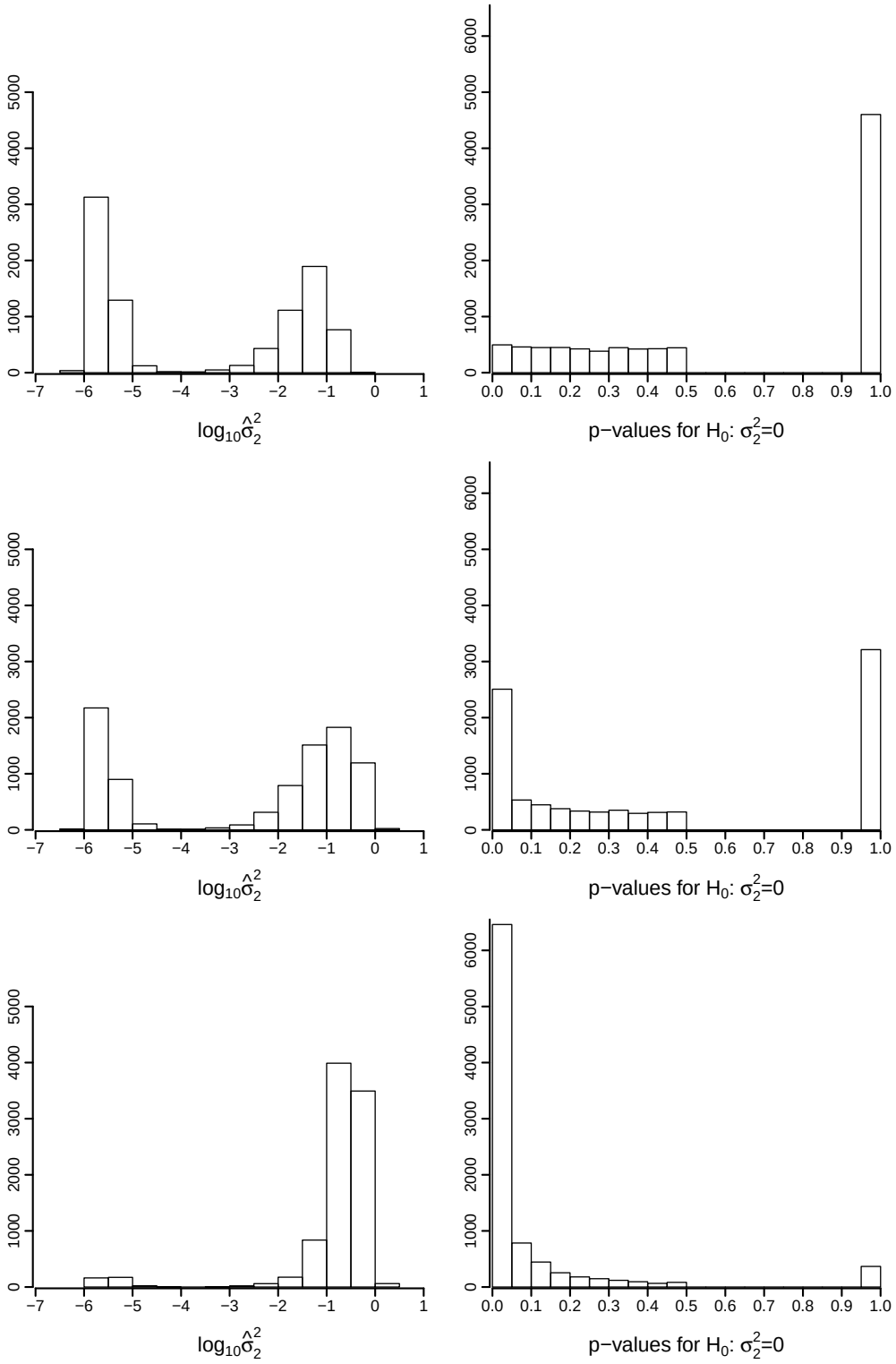


Figure 3.2: Estimates of blending error variance for simulated data. Log-estimates of the blending error variance σ_2^2 (left) and p-values (right) of RLRT ($H_0 : \sigma_2^2 = 0$) for simulated data. Top: perfectly blended individuals were simulated. The p-values of the interval [0,0.5) are uniformly distributed and nearly half of the transcripts have a p-value of 1. Middle: 3000 out of 9000 transcripts affected by imperfect blending of individuals. Bottom: all transcripts were simulated with imperfect blending.

to bottom) show an increasing proportion of large values, in full accordance with the increase in the simulated proportions of transcripts with a relevant blending error variance (which was 0, 1/3 and 1). The corresponding p-values (right panels of Figure 3.2, top to bottom) fairly mirror the same trend. The estimates for $\hat{\pi}_1$ approximated the simulated proportions of affected transcripts well. However, when it

Number or proportion of transcripts	Data set						
	Simulated			Experimental			
	s1	s2	s3	mouse	rat	bee	human
total	9000	9000	9000	18646	15923	14400	43256
crit. > 8	9000	9000	9000	8554	6264	13761	12348
sign VC	1	1794	6704	4329	6	7093	0
$\hat{\pi}_1$	0.005	0.295	0.918	0.75	0.29	0.68	0.40

Table 3.2: Number of transcripts with non-zero blending error variance. Results of the residual likelihood ratio tests of the hypothesis $H_0 : \sigma_2^2 = 0$ for transcripts exceeding the minimum expression level (crit. > 8). Numbers of transcripts with a significant variance component for imperfect blending (sign VC) were counted according to the FDR correction level of 5%. $\hat{\pi}_1$ is the estimated proportion of transcripts with $\sigma_2^2 > 0$.

came to the identification of individual transcripts, their number clearly lagged behind the proportions present in the data. Corresponding results are shown in Table 3.2. Differences in both models' abilities to find differential expression in the simu-

Number of transcripts identified	Data Set					
	Simulated			Experimental		
	s1	s2	s3	mouse	rat	human
m1 & m2	3112	3119	3128	3344	1636	787
m1	48	113	279	504	141	350
m2	4	13	29	516	12	154

Table 3.3: Number of transcripts identified as differentially expressed at an FDR of 5% by data set and model. Simulated data sets s1, s2, and s3 refer to scenarios were none, one third, and all transcripts were associated with a non-zero blending error variance component. The number of transcripts identified with both models is indicated by m1 & m2, transcripts identified solely with the null model (m1) or the full model (m2) are shown in the second-to-last and the last line.

lated data sets were also observed (Table 3.3). The null model yielded an average of 3407 expressed transcripts declared as differentially expressed, compared to 3157

from the full model. The average shared number is 3128, but the 3000 simulated as differentially expressed in a total of 9000 transcripts was clearly outbid by both models. Figure 3.3 shows the average numbers of four sets of transcripts and their

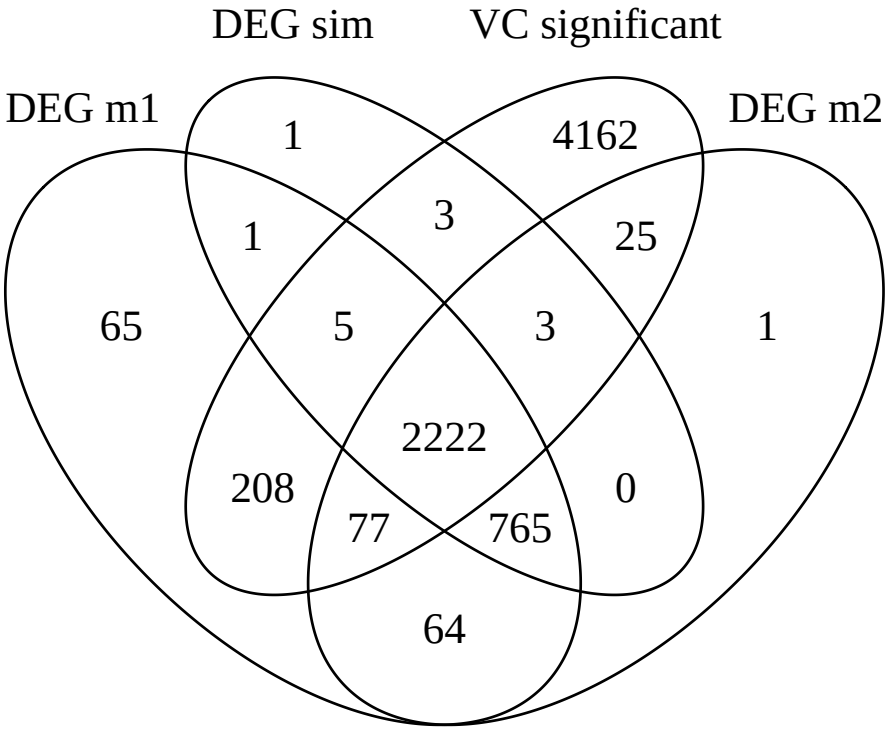


Figure 3.3: Sets of differentially expressed transcripts (DEGs) for both models and coincidences of transcripts with a significant variance component for imperfect blending. These were averaged over 100 repetitions of the simulated experiment based on the mouse design and variance components $\sigma_t^2 = 0.017, \sigma_b^2 = 0.094$ and $\sigma_z^2 = 2.7$ (all transcripts with effects for imperfect blending). The average counts of the sets of differentially expressed transcripts are labeled with 'DEG m1' for the null model, 'DEG m2' for the full model, 'VC significant' for transcripts with a significant blending error variance, and 'DEG sim' for the transcripts simulated as differentially expressed.

intersections: the set of transcripts with a simulated differential expression, one set of transcripts identified as differentially expressed for each of both models, and the set of transcripts, which were identified as connected with an attributable (larger than zero in terms of FDR) blending error variance. Upon counting the numbers in the intersection regions which corresponded to true discoveries, a similarly high power for both models was observed. Only 7 (m1) and 10 (m2) of the transcripts simulated as differentially expressed have not been found. But, adding the numbers

which correspond to false discoveries yielded a value of (1+25+64+77)/6000 = 0.028 for m2 and (65 + 208 + 64 + 77)/6000 = 0.069 for m1. This is clearly larger than 5%, the chosen level of permitted false discoveries. The number of transcripts incorrectly labeled as differentially expressed in the group of transcripts with a significant blending error variance was inflated by a factor of about three for m1 (285) in comparison with m2 (102).

Furthermore, in a series of simulations, the pooling technical variance σ_z^2 was varied within the range of $(0, 2.7]$. A plot of the obtained estimates of σ_2^2 against the simulated values $\sigma_z^2(e^{\sigma_b^2} - 1)$ (see section 7.1) shows nearly perfect consistency. The exception is some upward bias for very small simulated values, which can be attributed to the well-known properties of the REML-method (Swallow and Monahan, 1984).

Therefore, it can be concluded at the very least that tests for differential expression with the m1 model tend to be too optimistic, depending on the given experimental conditions. To summarize, should the model contain the additional random effect of imperfect blending, the statistical analysis yields results which agree very well with the simulated characteristics.

3.3.2 Experimental data

Histograms of log-transformed estimations of the variance components due to imperfect blending are shown in Figure 3.4. Estimates range from nearly zero (10^{-6}) to less than one hundred (10^2). A clear bimodal distribution can be observed in all cases, where the left part of each distribution (values less than approximately 10^{-3}) represents very small values close to zero while the other part represents more substantial values. In the mouse and the bee data, the proportion of transcripts with substantially large values clearly exceeds the proportion of small values. For the human data, the proportion of small estimates also prevails somewhat, while a balance between minor and substantial values can be observed for the rat data. This

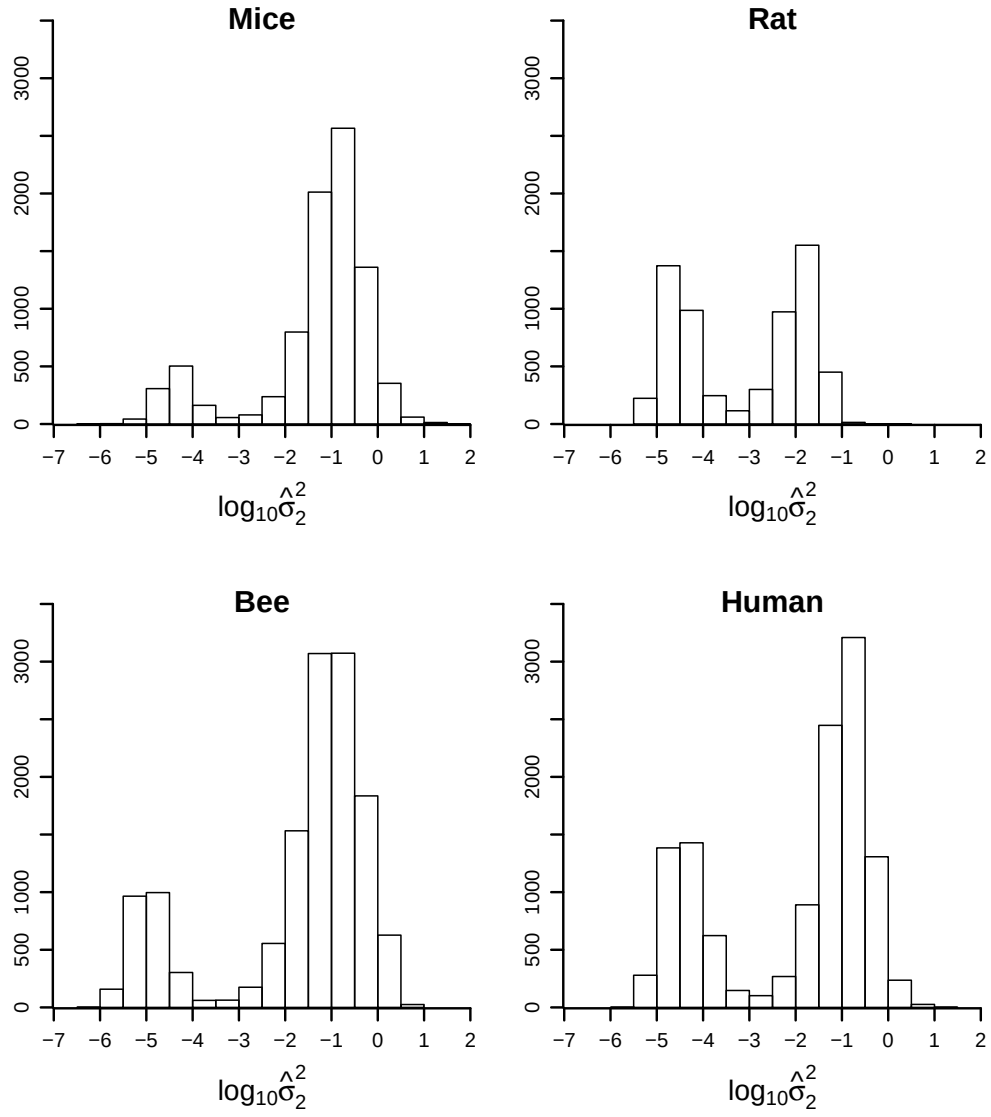


Figure 3.4: Histogram of log-estimates of the variance component σ_2^2 for the experimental data sets mouse, rat, bee, and human.

is also reflected in the average (over all transcripts) of all three variance components obtained with the reduced (m1) and the full (m2) models, as shown in Table 3.4. In light of the averages, the inclusion of a blending error variance had the consequence of a more or less reduced residual variance, most pronounced in the mouse and honeybee data. In the human data, the average residual variance remained almost constant, yet the average biological variance decreased - a phenomenon not observed in the other data sets. Distributions of the size of σ_2^2 relative to the total variance of a standard observation — with respective pool sizes of 5, 3, 4, and 3 for

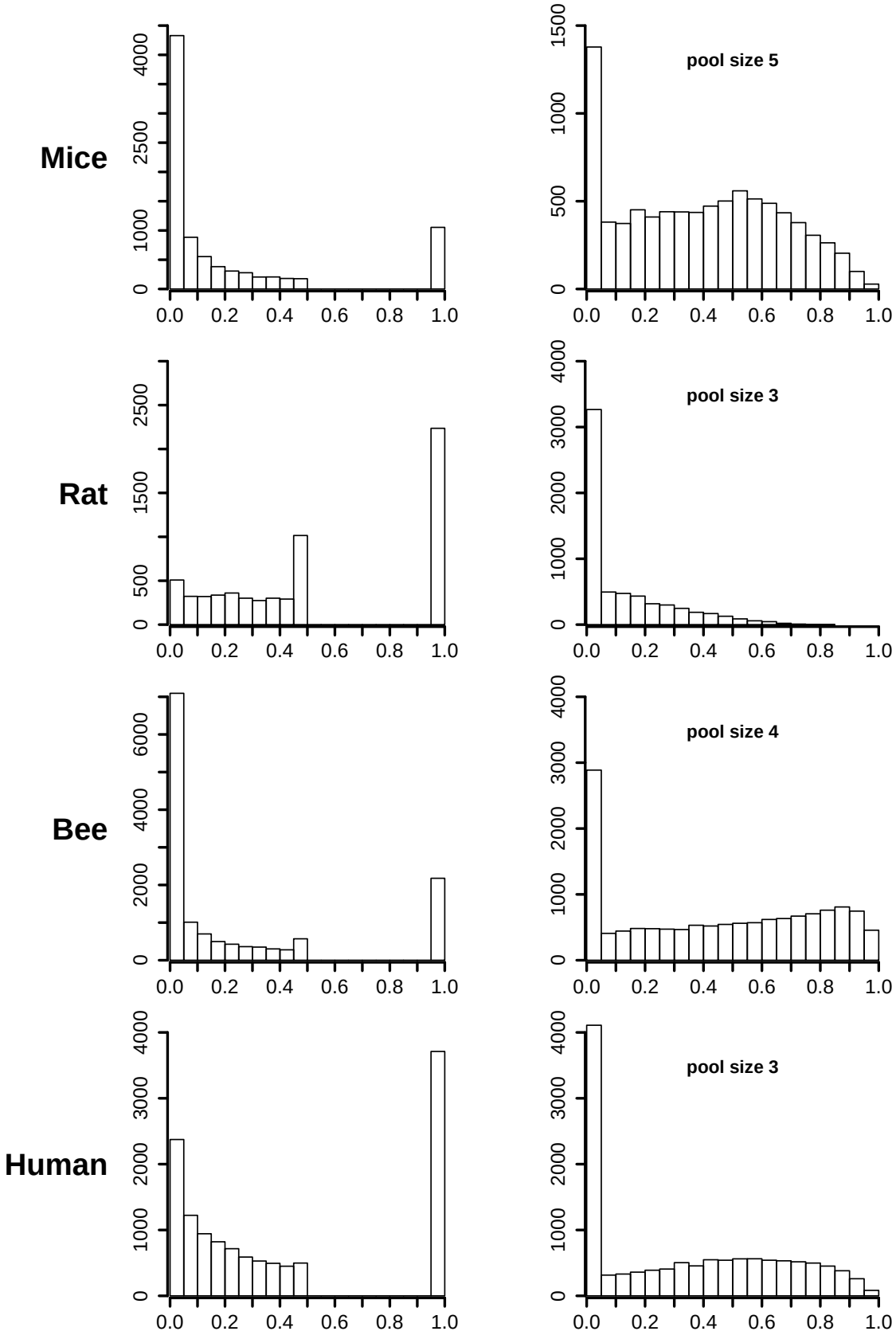


Figure 3.5: Diagram of p-values of RLRT and variance ratios. For each experimental data set, a histogram of p-values of the likelihood ratio test statistic for the test of $H_0 : \sigma_2^2 = 0$ are shown (left), as well as histograms of the variance components for imperfect blending, expressed as the proportion of the total variance (right) of a standard observation. y-axis: count of transcripts.

Mean estimated variance component	Experimental data set / model used							
	Mouse		Rat		Bee		Human	
	m1	m2	m1	m2	m1	m2	m1	m2
σ_e^2	0.037	0.017	0.010	0.009	0.104	0.035	0.062	0.060
σ_1^2	0.109	0.109	0.024	0.024	0.031	0.033	0.105	0.055
σ_2^2	-	0.295	-	0.011	-	0.215	-	0.155

Table 3.4: Mean estimated variance components. Estimated variance components for residuals (σ_e^2), biological effects (σ_1^2), and imperfect blending (σ_2^2) - averaged over all analyzed transcripts for the null model (m1) and the full model (m2).

mouse, rat, bee and human data — are given in Figure 3.5 (right, top to bottom). All distributions exhibit a clear spike near zero, followed by estimates that nearly exceed the full range of variance ratios. The rat data are an exception; hardly any values larger than 0.6 were observed.

These impressions are mirrored by the distributions of p-values from RLRT-tests for the hypothesis of a non-existing ($\sigma_2^2 = 0$) blending error variance (left panels in Figure 3.5, top to bottom). The number of individual transcripts, which could be associated with a non-zero blending error variance at a false discovery rate of 5%, varied strongly between data sets. There were 4329 of such transcripts in the mouse data and 7093 in the honeybee data, while only 6 were identified in the rat data and none at all in the human data (Table 3.2). These high numbers are consistent with considerable estimates for the fraction ($\hat{\pi}_1$) of non-zero variances in mouse ($\hat{\pi}_1 = 0.75$) and honeybee ($\hat{\pi}_1 = 0.68$) data (Table 3.2). Note that the respective estimated proportions were $\hat{\pi}_1 = 0.29$ and $\hat{\pi}_1 = 0.40$ in the rat and human data (Table 3.2), also indicating the existence of non-zero blending error variances in these two data sets, though almost no particular non-zero variance could have successfully been identified at the chosen false discovery rate of 5%.

Counts of differentially expressed transcripts detected with both models are shown in Table 3.3. About half of all transcripts analyzed were declared differentially expressed in the mouse data. About five hundred were exclusively detected with one of both models: 504 with the null model and 516 with the full model. The list of

the top 100 transcripts - ranked by their p-values - showed a large dissimilarity as indicated by a value of 0.11 for Kendall's correlation test. In the rat data, 1636 differentially expressed transcripts were jointly identified by both models, while 141 were solely found with the help of m1 and 12 with m2. No numbers appear in Table 3.3 for the honeybee data, as no differentially expressed transcripts were found. Finally, there were 1137 differentially expressed transcripts from the null model in the human data, from which only 787 were 'confirmed' by the full model.

3.4 Conclusions

In light of the large numbers of blending error variances diagnosed as greater than zero in the mouse and honeybee data, the practical relevance of this second kind of technical error has been clearly demonstrated. In both other data sets, estimates of $\hat{\pi}_1$, the proportion of positive blending error variances, may be taken as an indicator of their existence, though hardly any particular values could be identified, presumably due to a lack of power. As demonstrated mainly by simulation, there are also consequences for the detection of differentially expressed transcripts, in which the nominal FDR-level was shown to be too optimistic when the blending error variance was not taken into account. Therefore, we strongly recommend the application of adequate models (as described in Rudolf *et al.* (2013)) including random blending effects and their variances when observations from pools of different sizes are to be jointly analyzed.

4 Are there biomarkers for hygienic behavior of individual Apis mellifera workers?

4.1 Introduction

Colonies of *Apis mellifera* are threatened by the parasitic mite *Varroa destructor*. Presence of mites in a colony damages the population, since infested larvae hatch weakened, bees show degenerated bodies, and have a shortened life span (e.g. Boecking and Spivak, 1999). Chemical treatment led to problems including adverse effects on bees, contanimation of honey, development of resistance of the mites (e.g. Rosenkranz *et al.*, 2010; Rademacher and Harz, 2006).

Meanwhile there is consensus that breeding of resistant colonies is more sustainable and more efficient than drug treatment alone. Hygienic behavior of *A. mellifera* towards parasitized brood is a major part of this resistance and a target in selection (Büchler *et al.*, 2010). An enhanced rate of removal of diseased brood may restrict the effective reproduction of mites and free comb space for new larvae. As a result such colonies have a lower disease load, since mites are also vectors of diseases. The probability of colony collapse is reduced, which is mainly triggered by an overspill of the mite population (Genersch *et al.*, 2010).

Various measures have been proposed for operationalization a colony's ability to

control its mite population (e.g. Villa *et al.*, 2009; Rinderer *et al.*, 2014). Quantification of general hygienic behavior is often done on a special area on the comb. Brood is either severly damaged by applying an array of needles (pin test) or killed by deep-freezing. In the latter case a section of deep-frozen brood of standardized size is inserted and degree of brood removal is measured (e.g.) 48 hours later. In the study of Spivak and Reuter (2001) colonies that removed the freeze-killed brood from the comb section within 48 hours on two trials were considered as hygienic.
Varroa sensitive hygiene (VSH) is assessed as the rate of removal of mite-infested pupae. The number of infested cells per 100 non-infested cells was used to test VSH bees for differences in resistance to the mite in Harbo and Harris (2009). The level of natural infection was measured and correlated to the removal rate. By selecting on VSH they managed to get siginificant improvements with respect to VSH in comparison to unselected commercial controls.
Some recent sudies have investigated aspects of hygienic behavior at the molecular level. *Varroa* resistant colonies, that survived without treatment for several years, were compared with highly parasitized colonies by Navajas *et al.* (2008). They found 116 differentially expressed genes, a large proportion of them responsible for development of the nervous system and neuron excitability. Another study showed that the hygienic genotype of the social partners in the colony affect the hygienic behavioral performance and the brain gene expression of single worker bees (Gempe *et al.*, 2012), suggesting a complex genetic architecture underlying the hygienic output at the colony level. Quantitative trait loci (QTL) for individual hygienic behavior towards artificially killed brood were mapped by Oxley *et al.* (2010). Engagement of hygienic behavior was related to three QTL, which accounted for 30% of the phenotypic variability. By interval mapping Tsuruda *et al.* (2012) found a chromosome-wide significant QTL on cromosome 9 for the individual VSH trait, which was considered as perforating the wax capping, enlarging the opening of an already perforated cap, or removing a pupae. Workers were sampled if they engaged in these behaviors for at least two minutes, and if the targeted brood cell was in-

fested by *V. destructor*. The transcriptoms of antennae from VSH bees (observed performing VSH-related tasks) and non-VSH bees were compared in Mondet *et al.* (2015) by RNA-sequencing. Differentially expressed were 258 genes, and gene ontology analysis revealed that a category called 'defense response' was enriched in the VSH bee antennae.

Reliable molecular markers would probably be a great help in breeding for Varroa resistance, because it might supplement or even replace measures on the phenotypical level. Molecular markers can be used as surrogates for complex phenotypes, which are difficult to associate with a few genetic markers. Expression-based biomarkers are close to the phenotype, in fact they can be interpreted as a molecular phenotype (Schudoma *et al.*, 2012). Therefore, the development of a biomarker, considered as a set of genes with satisfactory discriminatory power whose joint expression pattern is predictive of class membership (Dziuda, 2010), is a goal in the investigation of hygienic behavior in the genome.

In this article, the steps of a search for a molecular marker for individual uncapping and removal behavior are presented, using data from a gene expression experiment with two-color arrays. Usefulness and further steps of research for breeding of colonies resistant to Varroa destructor is discussed.

4.2 Material and Methods

In the following the collection of bees and the design of the gene expression experiment are explained. Then, the preprocessing step of two-color array data with a linear mixed is described, and the applied re-sampling scheme for the biomarker search using logistic lasso regression.

4.2.1 Collection of hygienic workers and controls

In the behavioral experiment individual hygienic behavior of worker bees against Varroa-parasitized brood was evaluated. The worker bees were progenies from 7 young queen bees, which were bred from different queens of a selection line mated with unselected drones, and were back-crossed with drones from different queens of the selection line. Rearing of bees was standardized in the same hive. At the age of one day bees were individually marked. 2000 worker bees from 7 colonies participated in the behavioral experiment of 48 hours duration. This population of bees from 7 different genotypes (of queen bee, also referred as colonies) were presented a honeycomb which contained a section with Varroa-parasitized brood for infrared video observations. After 48 hours they were simultaneously killed in liquid nitrogen for age-specific gene expression.

After video analysis of 7 repeats (runs) of this setting, worker bees, which engaged the opening of at least one diseased cell, and involved herself in at least one removal, were selected into the hygienic behavioral class. Worker bees from the same run and colony with no activities were sampled as controls.

4.2.2 Design of the gene expression experiment

For the gene expression experiment the two-color array (TCA) platform was used, the honeybee whole-genome oligonucleotide microarray representing 13440 transcripts. In the 7 repeats of this setting worker bees of 4 colonies were observed showing hygienic behavior, in total 43 bees. As not in every possible of the $7 \times 4 = 28$ groups of bees appeared a hygienic bee, a dye balance design (Knapen *et al.*, 2009) was sequentially applied, which means that the hygienic bees of each group had to be labeled equally often with red and green dye. By doing so, the difference of the effects for behavior is separated from dye effects, as well the effects of colony and experimental run. In groups with an odd number of hygienic bees therefore one of them had to be used a second time. On that array a sample of a new control bee

of the same group was hybridized. As the number of odd groups was nine, the 43 hygienic bees were matched with 52 controls on 52 arrays. Another 12 two-color arrays (TCA) were labeled with mRNA samples of hygienic bees only for a comparison between colonies. This design, developed for high power to detect differences in the transcriptome between the behavioral classes, created data with correlated measurements. The design was planned for three different nerval tissues: Antennae, anntenal lobe and mushroom bodies. Supposing that a successful detection of *Varroa*-parasitized brood relies on the olfactory sensitivity, whose recognition is connected to mushroom bodies, their RNA was examined for a biomarker search.

4.2.3 Classification with two-color array data

In the classification context a defined class membership per sample (bee) is needed for usual multivariate methods. This means that a proper data matrix has to be used, which contains for each transcript a vector (i.e. column) of gene expression values of the samples.

Usually, data from TCA experiments are prepared for analyses by normalization, consisting of background correction, adjustment of distribution and logarithmic transformation. Afterwards it consists of differences from measurements of samples, in detail M-values as log-spot-ratios of intensity measurements of red versus green channel. Because of this difference of log-intensities of red and green channel, setting up a gene expression matrix is a non-standard procedure.

In the simple case of a reference design (several samples compared with a unique reference sample), the gene expression matrix can be built by these M-values. Otherwise a preprocessing step for preparing of the gene expression matrix is needed. A few examples for biomarker search in this case exist. In an experiment described in Nota *et al.* (2010), log-spot-intensities (A) were used to back-transform to single channel log-intensities $R = A + \frac{1}{2}M$ and $G = A - \frac{1}{2}M$) of the individual samples. Applying this procedure, it seemed to impact correlations in our data matrix, and

performance of classification methods in terms of prediction errors were poor.
Furthermore, as the data were produced from an experimental design including replications, it is not obvious how to handle multiple measurements, as averaging changes the variance and a (random) selection of a single sample data point is arbitrary. Thus, there was need to develop an new approach.

4.2.4 A mixed model approach for preprocessing

The particular structure of the data, which consists of differences of log-intensities with correlations and multiple used samples, were accounted for by a mixed model based data preparation step. The summary of all available information for each sample was carried out by the estimation of a random bee effect, which was then entered as a phenotypic value in the data matrix. By doing so, one accounts for the variance structure. No matter an individual is involved in one or more measurements, there is only one solution for each bee.
Let $\mathbf{M}_{n_a}$, with n_a the number of arrays, be the vector of differential log-expression values for a transcript j, where $j = 1, \ldots, p$ with p the total number of transcripts (features).

$$\mathbf{M}_{n_a} = \mu + \mathbf{Z}\mathbf{x}_n + \mathbf{e}_{n_a} \tag{4.1}$$

The vector $\mathbf{M}$ is modeled via equation 4.1 with a mean effect μ, which contains also the distortions through color, since for all arrays the log-values of the green channel are subtracted from the red. The bees used in the hybridization design are numbered $i, \quad i = 1, \ldots, n$. A random bee effect $x_n \sim \mathbf{N}(0, \sigma_x^2)$ is included. Then $\mathbf{Z}$ is of dimension $n_a \times n$ and contains two non-zero entries per row, 1 for the individual used for the red channel of the array and -1 for green. Residual errors in $\mathbf{e}_{n_a}$ are assumed to contain also distortions due to measurement.
Thus, the distortions through channel effects and measurement are excluded. The fitted random bee effects $\mathbf{x}_n$ contain information about behavior and colony and other biological effects. The gene expression matrix $\mathbf{X}$ of dimension "samples x

transcripts“ is built column by column with the fitted random effects (scaled by σ_x) from the above linear mixed model.

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \tag{4.2}$$

Of course, so far no class information has been used. With this gene expression matrix and the vector of class membership biomarker search can now be started.

4.2.5 Adaptive Lasso for binomial data

For the selection of features (chose a few transcripts out of several thousand) in a sparsity situation, with only a few truly relevant predictions within a large number of candidates, a penalized least squares method is well suited (Hastie *et al.*, 2009). The lasso is a regularization technique for simultaneous variable selection and estimation of predictive coefficients in a linear model (Tibshirani, 1996). Zou (2006) proposed the adaptive lasso to fulfill consistency in variable selection. Further they provided a method for generalized linear models, which is applied here for binomial data. Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ be the response and consider $\mathbf{x}_1, \ldots, \mathbf{x}_p$ the columns of the gene expression matrix $\mathbf{X}$

$$\hat{\boldsymbol{\beta}}^{*(n)} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(-y_i(\mathbf{x}_i^\top \boldsymbol{\beta}) + \log(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}) \right) + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left| \beta_j \right|, \tag{4.3}$$

where λ is the nonnegative regularization parameter and $\boldsymbol{\beta}$ the vector of coefficients $\beta_j, \quad j = 1, \ldots, p$. In contrast to the common lasso, weights $\hat{w}_j$ are inserted for each coefficient - to be estimated in an initial step. The adaptive lasso estimates can be found by the LARS algorithm Efron *et al.* (2004). Inside the open-source statistical programming package R (R Core Team, 2012), the adaptive lasso implementation of Kraemer and Schaefer (2012) was used for calculations.

4.2.6 Course of search

For about 2000 worker bees from 7 colonies, only 43 bees from 4 colonies were identified as showing hygienic behavior. After normalization of the two-color microarray

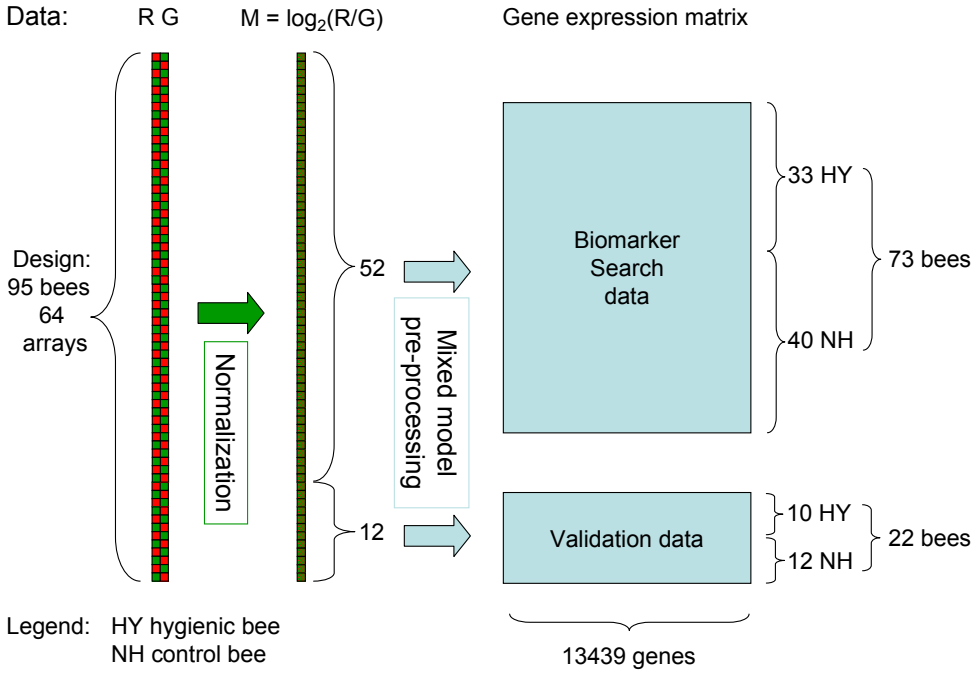


Figure 4.1: Scheme of data generation and transformation to gene expression matrix of search data and validation data.

data, 64 arrays with 13439 transcripts each were available. Therefore the gene expression matrix would have dimension 95×13439 in maximum. For an overview how the data were processed see Figure 4.1. For the sake of evaluating the biomarker candidate, 12 arrays were removed before the search. As a validation sample, hygienic and control bees of each of the 4 colonies were chosen, to be adequate for the investigated data space. This so-called internal validation was done for two main reasons. First it is considered to provide valid assessments of the model performance. Secondly, external validation data is difficult to access for a model building researcher and might, if not representative, even produce misleading results (Kessler, 2007).

The mixed model approach introduced above was then separately applied to both, the search data and the validation data. All splits of the data were done with respect

to the ratio of hygienic and control bees. That means, the validation data consisted of data points from 10 hygienic and 12 control bees.

For the biomarker search the following procedure was repeated 5000 times (re-sampling). The generated search data set of dimension 73×13439 was randomly split into training set and test set (about 2/3 to 1/3) containing 49 (22 HY to 27 NH) and 24 (11 HY to 13 NH) bees also stratified by the behavioral class. With adaptive lasso a selection of transcripts and estimation of coefficients (β^*) was executed in the training set. Coefficients β_i^* of non selected transcripts were set to zero. With these coefficients a value for each sample of the test set was calculated as

$$\frac{\exp \sum_{i=0}^{p} x_i \beta_i^*}{1 + \exp \sum_{i=0}^{p} x_i \beta_i^*} \quad .$$

Due to the logit transformation these values are inside the interval (0,1) and can be interpreted as susceptibility of a bee to show the hygienic behavior. By default, the cutoff for classification was set to 0.5, and the prediction error as proportion of wrongly classified samples was determined. Also statistics of used features were generated, including number of appearances. The most used features were then used to build a candidate biomarker.

4.3 Results

The results in terms of prediction error and selected features are presented and used to derive a set of transcripts for a biomarker candidate, which was then tested in an internal validation. The learning took place with the 73 bees search data and the classification in an internal validation with the 22 bees validation data.

4.3.1 Search for biomarker with adaptive lasso based on resampling

In Figure 4.2 the histogram of the prediction errors from 5000 resampling steps is shown for test set size 24 (around one third of the training data set). The re-

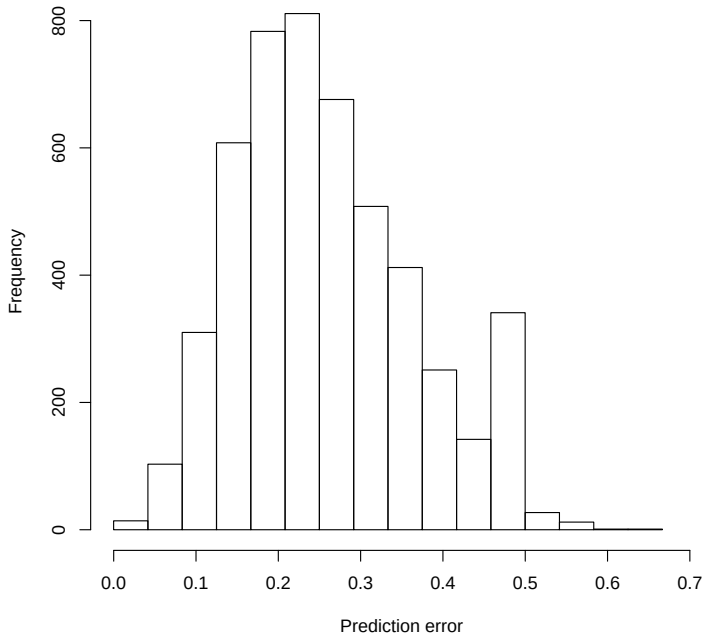


Figure 4.2: Histogram of prediction errors from a resampling with 5000 repetitions. In each run 49 from 73 bees of the search data set, were used to train a classifier with adaptive lasso. Prediction errors were then determined as wrongly classified cases of the remaining 24 bees of the test sets.

sulting distribution of the prediction error is appropriate to assess the ability of the biomarker to classify new samples (Dziuda, 2010). It can be characterized by the median $\frac{5}{24}$ and interquartile range of $\frac{1}{8}$. A more detailed view on the results of adaptive lasso in the resampling procedure is presented in Figure 4.3. It shows the prediction error according to the numbers of transcripts used by adaptive lasso. Although in average 13 transcripts were used, in 254 from 5000 runs no feature could be selected and the error is 0.458. This number is the ratio of the number of hygienic bees (11) in a test set of 24, if stratified sampling with respect to 43 hygienic and 52 control bees is applied. Already with 5 used transcripts the prediction error can be expected

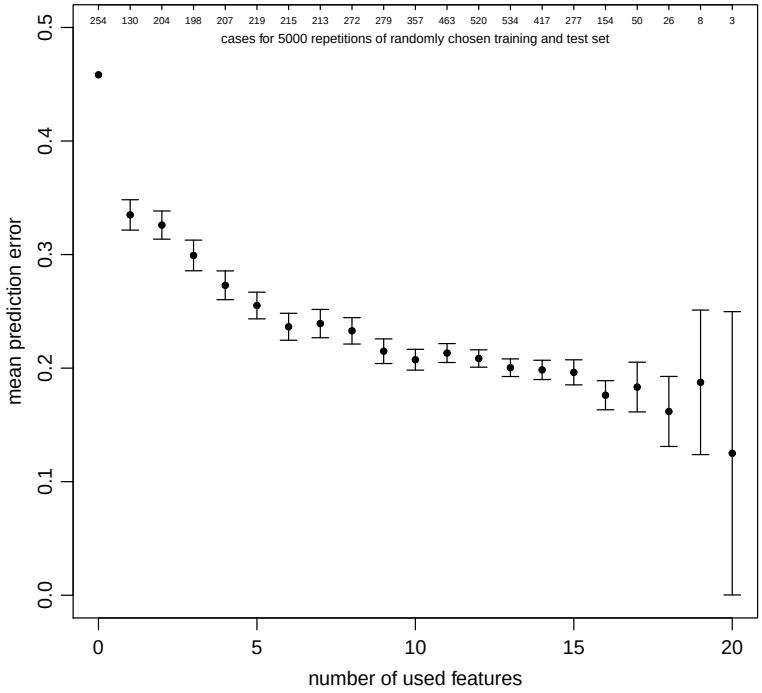


Figure 4.3: Histogram of prediction errors by number of used transcripts. Confidence intervals were calculated from standard errors of prediction errors of runs which were equal according to the number of selected transcripts.

to lie around 25%, with further growing number of transcripts small improvements can be seen (prediction error around 20%). To determine the biomarker candidate the selected transcripts in every run were stored and their frequencies are shown in Figure 4.4. The first 4 transcripts had a clearly higher frequency then the following transcripts, whose numbers of usage differ not very much. Therefore the size of the biomarker candidate was set to 4, including the transcripts AM03409, AM09219, AM07292 and AM01976 (oligo IDs). From a logistic model fit for the search data set of 73 bees the coefficients for the top 4 ranked transcripts were determined and are displayed in Table 4.1.

Transcript	(Intercept)	AM03409	AM09219	AM07292	AM01976
Coefficient	-0.2412685	0.3511680	-0.8159310	-0.3421664	-0.3893987

Table 4.1: Coefficients of logistic regression with the biomarker candidate of 4 transcripts. The logistic regression model was fitted with all 73 bees of the search data set.

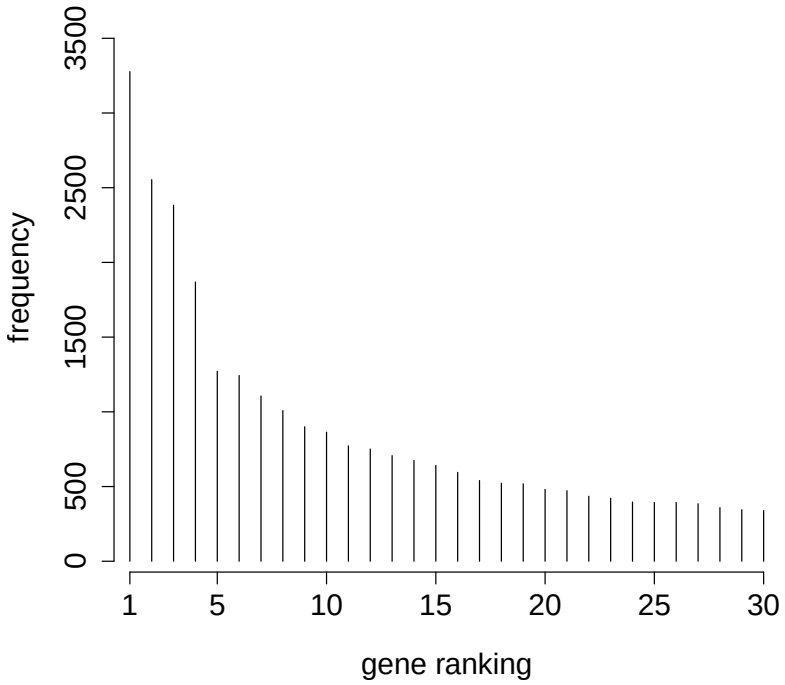


Figure 4.4: Counts of most used transcripts, which were selected by adaptive lasso in 5000 repetitions.

4.3.2 Predictions for bees from validation data

This biomarker candidate set was then applied to classify the bees from the 12 arrays, which were not used while learning. From a new mixed model fit, the data points of the validation data were calculated. Then the bee effects were used together with the coefficients from the learning to calculate the predictions. Denoting the biomarker coefficients from Table 4.1 with β^* and the respective column of the validation data matrix to x, the predictions

$$\frac{\exp \sum_{i=0}^{4} x_i \beta_i^*}{1 + \exp \sum_{i=0}^{4} x_i \beta_i^*}, \tag{4.4}$$

with $x_0 = 1$ were used to classify the 22 cases at a cutoff of 0.5. It turned out, that 6 bees from 10 hygienic were correctly classified and 10 out of 12 controls. This gives a sensitivity of 60%, specifity of 83% and a prediction error of 27% for the validation

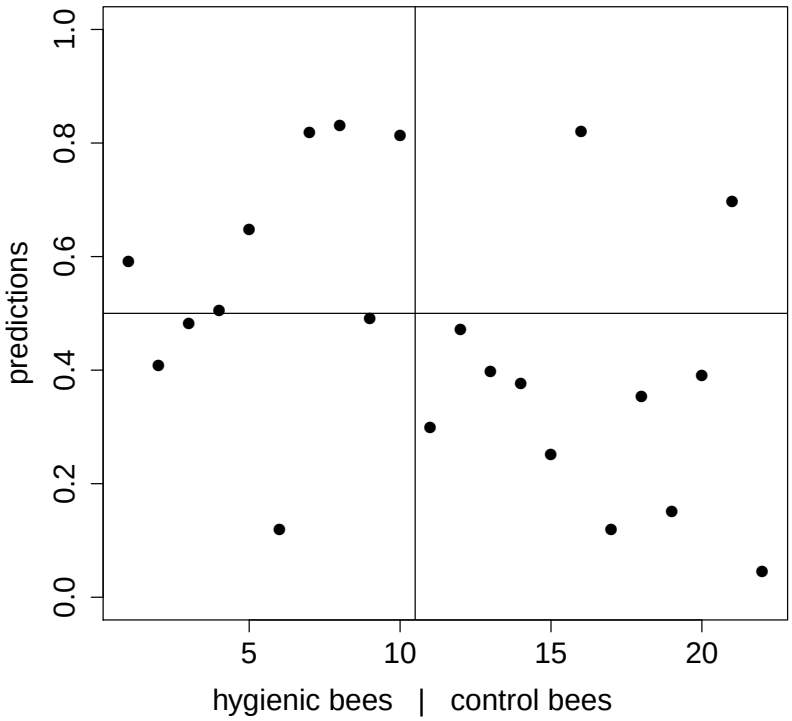


Figure 4.5: Predictions for 10 hygienic and 12 control bees from the validation data. The bees were grouped by their true behavior, left side hygienic and right side control bees. The predictions were calculated according to Equation (4.4) with the coefficients from Table 4.1 and a cutoff of 0.5. Thus, the upper left and lower right quadrant of the figure contain the correctly classified samples.

data.

4.4 Discussion

4.4.1 Mixed model preprocessing

Two-color array data was transformed for biomarker search with an adapted approach. For each bee a single data point was sequentially determined by fitting a mixed model for every transcript.

For multivariate methods a conservation of the correlation structure while transforming the data is essential. It turned out, that the correlation structure of transcripts is

only little affected by the transformation proposed opposite to back-transformation to single channel data. This was investigated by calculating correlations of tran-

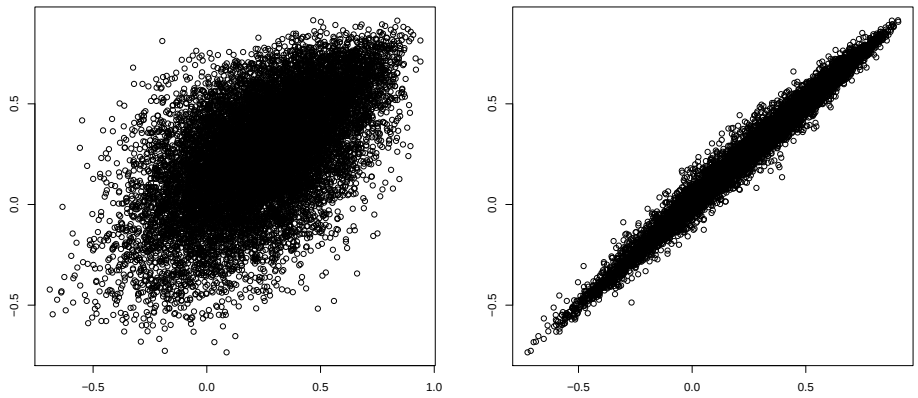


Figure 4.6: Scatterplot of correlations between consecutive transcripts for **M**-vectors (y-axes) before transformation (array data) and for sample vectors (x-axes) after transformation (GEM data of dimensions 95×13439). On the left the sample vectors (as columns of GEM) were derived through back transformation to red and green channel values (128) and crossing out rows until unique bees (95). On the right, the estimated random bee effects from the proposed preprocessing were entered into the columns transcript-wise.

scripts for the data matrix of **M** (dimensions: 64×13439) and compare with correlations after transformation in the sample data matrix (dimensions:95×13439). The latter were determined a) by back-transformation two red and green channel data (128 bees) and crossing out rows until each bee is represented only once (left) and b) with the proposed preprocessing which uses random bee effects (right). As not all $13439 * 13438/2$ correlations could be displayed, they were calculated for consecutive transcripts (1 vs. 2, 2 vs. $3, \ldots$, and 13438 vs. 13439). Then corresponding pairs (before and after transformation) of correlations were used as coordinates of the spots in Figure 4.6. On the left side can be seen, that the structure changed due to back-transformation to single channel values. On the right, the spots are mostly very close to the diagonal, which indicates a very good preservation of the correlation structure.

This approach might be applied to a wide range of experiments with in principle arbitrary designs, which have to be considered in the design matrices of the mixed

model. The only assumption is that there need to be replications in the design, either repeated arrays or repeatedly used samples, for estimability of the variance component and random bee effects.

4.4.2 Biomarker candidate with 4 transcripts

Four transcripts from the top list of used transcripts from 5000 resamplings were selected. Based on homology searches the top ranked transcript AM03409 is possibly involved in neuron development. This pathway was also found to be upregulated in Varroa resistant bees by Navajas *et al.* (2008).

The results of the validation showed the ability of the biomarker candidate set to classify unknown bees according to hygienic behavior. The prediction error of 27% for the validation data was above the median prediction error from Figure 4.3. This deviation might be connected with the low number of bees in the validation data. The specifity was even higher than the mean during the learning. For reinforcement of the sensitivity from a level of 60% a cutoff value lower than 0.5 might be considered.

An independent validation study requires testing of many colonies. The data in the HyBee project was collected in an elaborate experiment with visual identification of activities of bees. This high efforts would have to be repeated for a validation study, as known class memberships are inevitable. For such a study a change of the data generating platform should be considered, because direct measurement for instance with RT-PCR are known to be more precisely than microarray measurements. In any case, using the derived candidate biomarker shall follow the same recipe for the independent validation data. It suffices to generate expression values for the biomarker transcripts, and on the resulting gene expression matrix of small dimension the calculated coefficients of the logistic regression procedure can be applied. The candidate biomarker was chosen with 4 transcripts, because of the ranking in usage. Also an increase in the number of transcripts in the biomarker would be

possible, by adjusting to this number before the logistic regression for estimation of coefficients takes place.

4.4.3 Outlook on possible applications

The data here was collected age-specific, as the bees were killed simultaneously with liquid nitrogen. This is impossible in a practical application for beekeepers. For economical reasons the evaluation of the hygienic potential of colonies might be done with randomly chosen bees. But, the proportion of hygienic bees in a colony is considered in the low one-digit percentage range, which causes an uncertainty how many of them are among a collection of say 50 bees per colony. If using pooled samples costs would be controlled, but methods have to be adapted. For prediction using pooled samples, also a biomoarker search with pooled samples is recommended. This biomarker is then not applicable for the prediction of single samples (Telaar *et al.*, 2013). Therefore, it is necessary to decide which biomarker, for pooled or single samples, is more useful for beekeeping.

A further possibility would be a pre-selection of small colonies. An RNA-marker could be used for testing young queen bees, via their workers, which shortens the generation interval and helps selection progress.

Furthermore, for assessment of colonies with a desirable low infection pressure a biomarker might help. Analyzing worker bees of colonies with a desirable low infection pressure, they might differ in their hygienic scores (value of biomarker averaged over bees). In the case of low scores, causes of low mite population might be a good (chemical) treatment or a low presence of mites in the region by chance, and with high scores the worker bees might be keeping the mite population low by early engaging hygienic behavior. In the latter case, such a colony would be valuable for breeding purposes.

The number of transcripts of a biomarker is important for the strategy of its application. For a few transcripts only evaluating with Real-Time-PCR would be preferred

for its higher precision, in the two-digits region an attempt of production of a bee-chip can be made.

There exist projects which studied breeding values of bees for brood care based on the Pin test or freeze-killed brood (e.g. Bienefeld *et al.* (2008)). Using a biomarker to assess hygienic behavior for such bees with a high breeding value, a relationship would be revealing.

After a successful validation study a possible application of the biomarker would be the analysis of colonies, where only some worker bees are collected and a minor number of transcripts is analyzed for assessing their hygienic potential. One advantage of this procedure is that bee combs need to be processed only once, instead of inserting a section of prepared brood and later counting success in removal, what is user-friendly for beekeeping.

Measurement of gene expression data may be a useful approach for the investigation of the regulation of traits in the transcriptome. The transcripts of a RNA marker could be used to derive a phenotypic value for single bees. This quantitative values (interpretable as probability to show hygienic behavior) can be used for breeding purposes and other analyses.

5 Summary and Discussion

The aim of this thesis was to study aspects of design and statistical analyses of gene expression experiments, initiated through the HyBee project (details in 1.2). In chapter 2 a new and more general design strategy for pooled gene expression experiments was developed. The bias through variable pool sizes was canceled by a design with a symmetrical structure of the pools between treatments. A general condition for unbiased contrasts was derived, which applies to one- or two-color array experiments. As a consequence, it became possible to include different numbers of individuals into pools of the same gene expression experiment. Therewith all 'special' individuals can be measured more efficiently than in a design with equally sized pools (more individuals while limiting the number and costs of the arrays).

The proposed linear mixed model for log intensities of gene expression is appropriate to account for the variance heterogeneity in analyses of experiments with variable pool sizes. Therein, an effect for blending of individual samples and the corresponding variance component was introduced, with the practical relevance checked in chapter 3. Experimental data sets from four different species were analyzed with the full model and with a reduced version, where this kind of technical error is lacking. The impact of neglecting the blending error variance on the type I error was shown in tests for differentially expressed transcripts.

In the third main part, chapter 4, an example of biomarker search was given in the context of hygienic behavior of honeybees. Furthermore, a novel method to generate the gene expression matrix was presented with respect to the correlation structure of the data. This method allows more experimental data sets to be used

in a biomarker search, namely from two-color array designs. A biomarker candidate with four transcripts was derived by using a known penalized regression approach in a re-sampling scheme. The search of a molecular marker will contribute in breeding *Varroa*-tolerable colonies.

5.1 Variable pool sizes

In the behavioral assays of HyBee, activities of bees versus the mite *V. destructor* were studied. Inside the comb, bees of seven colonies were presented with a diseased brood section. This experimental setting was repeated seven times for a duration of 48 hours each. The aim was to identify and analyze bees showing hygienic behavior. This was considered as the uncapping and removal of *Varroa*-parasitized brood cells. The aim of the microarray experiment was to analyze all cases, which exclusively executed both of these tasks. Due to the limited number of arrays, it was planned to use pooled samples, if too many cases for individual sample hybridizations had been found. Using a two-color array platform, direct comparisons of hygienic and control bees were possible. To reduce distortions, only bees of the same colony and runs should be hybridized on the same array. The number of cases showing hygienic behavior in a group (same colony and run) is not uniform. Therefore, a design with variable pool size (the number of individual samples blended into a pooled sample) was considered and conditions for adequate statistical analyses were derived.

Pooling designs and its impact on detection of differentially expressed genes have previously been investigated in several publications. Most of the work was done for a fixed number of individuals and a varying number of arrays according to a uniform pool size (e.g. Kendziorski *et al.* (2003)). There are major concerns about variable pool sizes in an experiment. It leads to biases in measurements depending on the size of pools and also to variance heterogeneity. In chapter 2, designs with flexible pooling were examined for tests of unbiased contrasts, by choosing the number and size of control pools adequately. Approximation formulas of Zhang *et al.* (2007) for

expectation and variance of pool-signals were used to quantify the bias (equation 2.3). The vector of biases $\mathbf{h}$ was introduced into a model for gene expression data. A generalized least squares estimator of biased measurements was calculated, as well as the expectation of an estimable contrast function according to the hypothesis that there was no difference between treatment effects. A condition could be derived to check if a contrast function can be estimated unbiasedly in the equation (2.8). It turns out that it suffices to oppose each pool of a treatment class with an equally sized pool of controls. It was also shown that this rule can be applied to a general case with more than two treatments. Furthermore it is also applicable for d multivariate measurements. They could appear through measurements of the same quantity of interest at different times or measurements which are distinguished by a second factor (like tissue from which RNA were extracted). Such designs require more effort, because for each additional factor the pool sizes have to be leveled.

The inhomogeneity in the measurement variance of experiments with variable pool sizes could be accounted for as follows. After decomposition of the pools variance and allocation of corresponding variance components (in equation 2.10), random effects were assigned for composition (u_1) and mixture (u_2) of pools in the linear mixed model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}$. The correlation matrices for the two random effects ($\mathbf{G}_1$ and $\mathbf{G}_2$) have entries, which are functions of pool size. Therewith, the modeled variance structure (equation 2.12) correctly considers the variance heterogeneity due to variable pool sizes. The variance parameter $\sigma = (\sigma_1^2, \sigma_2^2, \sigma_e^2)^\top$ can be estimated using restricted maximum likelihood (REML). Then, from the mixed model equations the solutions of fixed and random effects can be determined. Regarding the tests for differential expression, in general F-tests for the contrasts $\mathbf{K}^\top\hat{\boldsymbol{\beta}}$ have to be used. It should be noted that, for these tests of fixed effects in the mixed model, adjustments have to be done, including the covariance matrix of $\hat{\boldsymbol{\beta}}$ and estimation of degrees of freedom (Kenward and Roger, 1997).

The above system of one-color arrays can be applied to two-color arrays by building a matrix of differences $\mathbf{D}$, which has a row for each two-color array with the entries

1 and -1 for the two selected pools and zero otherwise. Then it follows that $\mathbf{Dh} = 0$ is a sufficient condition for unbiasedness of contrasts. The practical consequence is that only two pools of the same size are recommended to be hybridized together on an array.

With the theory presented in chapter 2, a general framework for modeling of gene expression experiments with pooled samples was developed. Therein, included as a special case are designs with individual samples, if pool size is one and the second variance component σ_2^2 is removed from the model.

In Zhang *et al.* (2007) simulations were tailored for the influence of pooling technical variance (σ_z^2) on the power to detect a certain mean class difference in pooled sample designs. Since a mistake led to the underestimation of the impact of σ_z^2, a few simulations were recalculated and provided in section 2.5. Thus, some loss of power (up to 4%) in the dependence of pooling technical variance could be shown. Former constrains of pooling experiments with variable pool size are now vanquished, since a uniform pool size in the whole experiment is no longer neccessary. The benefit is shown by an example of a fictive result of the behavioral experiment with three colonies in figure 5.1. Assuming that for every run one pool is available, with the flexible pooling design, 13 individuals can be analyzed. In the case of equally sized pools, only nine could be used. With more individuals power is expected to increase. The gain in power for such experiments was further investigated by simulations. In figure 5.2, results of a simulation are shown comparing the power to detect differential expression for three design strategies. Applying the flexible pool size design, the power to detect a mean class difference was higher than in the design with the equally sized pools. The vertical dotted lines indicate the number of pools to reach a power of 95%. It can be concluded that, with the flexible pool size design, less arrays are needed to get a certain power level.

Flexible pool size designs might be used in analyzing families with a high number of offspring. Individuals with a special property can be included in the design by the number in which they appear for a block of the experiment by increasing or

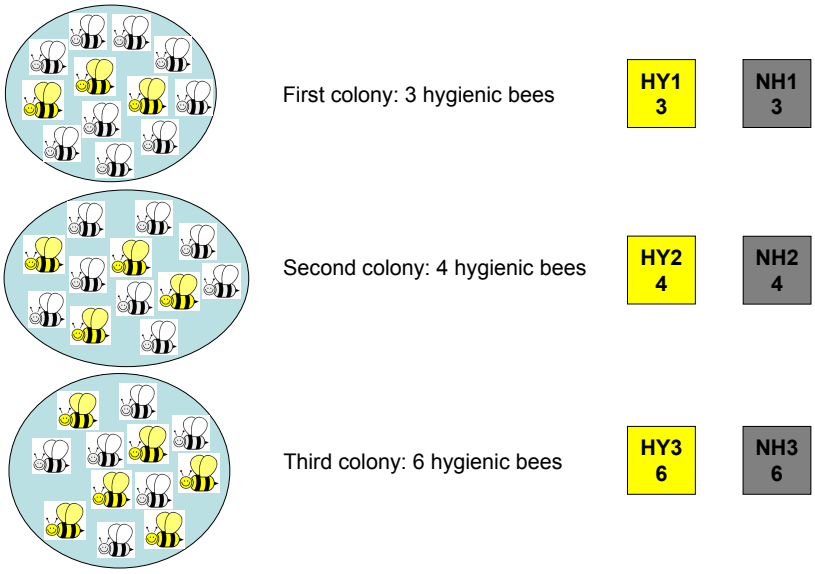


Figure 5.1: Design with variable pool sizes, which schematically shows different numbers of hygienic bees in a run (or a colony) and corresponding pool sizes. On the left side, symbolic colonies with different numbers of hygienic bees (yellow) and non-hygienic bees (gray) are displayed. Rectangles on the right side symbolize the pools for each run, with behavioral classes (HY: hygienic, NH: not hygienic) and pool size.

decreasing the pool size.

5.2 Inaccuracies due to blending

In chapter 2, a mixed model for log-transformed gene expression data from designs with pooled samples was developed; additionally, it considered blending samples by a random effect and a corresponding variance component. By applying pooling, RNA from different individuals sharing the same experimental conditions and explanatory variables are blended and their concentrations are jointly measured. As a matter of principle, individuals are represented in equal shares in each pool. However, some degree of disproportionality may arise from the limits of technical precision. As a consequence, a special kind of technical error occurs, which can be modeled by a respective variance component. The theory — allowing for variable pool sizes — has been applied to four microarray gene expression data sets from different species in order to assess the practical relevance of this type of technical error in terms of

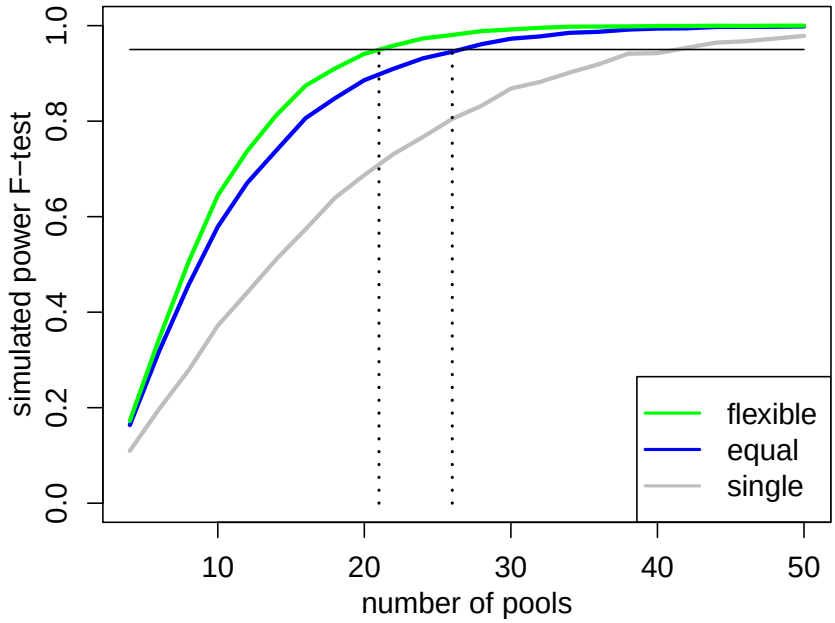


Figure 5.2: Comparison of power for three design strategies. Different number of runs (with two pools each, one pool for treatment and one for control) of an experimental setting with an expected number of five individuals per pool for a certain mean class difference $\Delta = 0.4$. The simulated variance components were $\sigma_b^2 = 0.08$,$\sigma_t^2 = 0.2^2$, and $\sigma_z^2 = 0.1^2$. The number of individuals of the treatment class was simulated from a Poisson distribution. Then, for design 'flexible', an equally sized pool was opposed for each run; for design 'equal', the minimum number was chosen as pool size; and for design single only two single samples were chosen.

significance and size of this variance component. Variable pool sizes have an effect on the variance of measurements. This was taken into account by considering how many individuals are allocated to each pool, and by introducing a random effect for blending with a corresponding variance component, called blending error variance. Experimental data sets from four species were chosen with the condition that the design included, firstly, pools of a different size, and, secondly, replications of measurements. This is necessary for the estimation of all desired variance components. The following linear mixed model (m2) was used for background-corrected and normalized log-intensities $\mathbf{y}$ of a certain transcript, where the length of the vector $\mathbf{y}$ equals the number of arrays: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}$. Random effects of single individuals were assumed to be independently identically distributed with biological variance σ_1^2, observations from a number of γ_i individuals have biological variance

$\frac{\sigma_1^2}{\gamma_i}$. The random effect for blending u_2 consists of one effect per mixture, the associated variance component is σ_2^2.

The starting point was a data set from mice, with the specialty, that for some pooled sample observations there also existed observations of individuals which were blended into these pools. In figure 5.3, it can be seen that the first 10 individuals

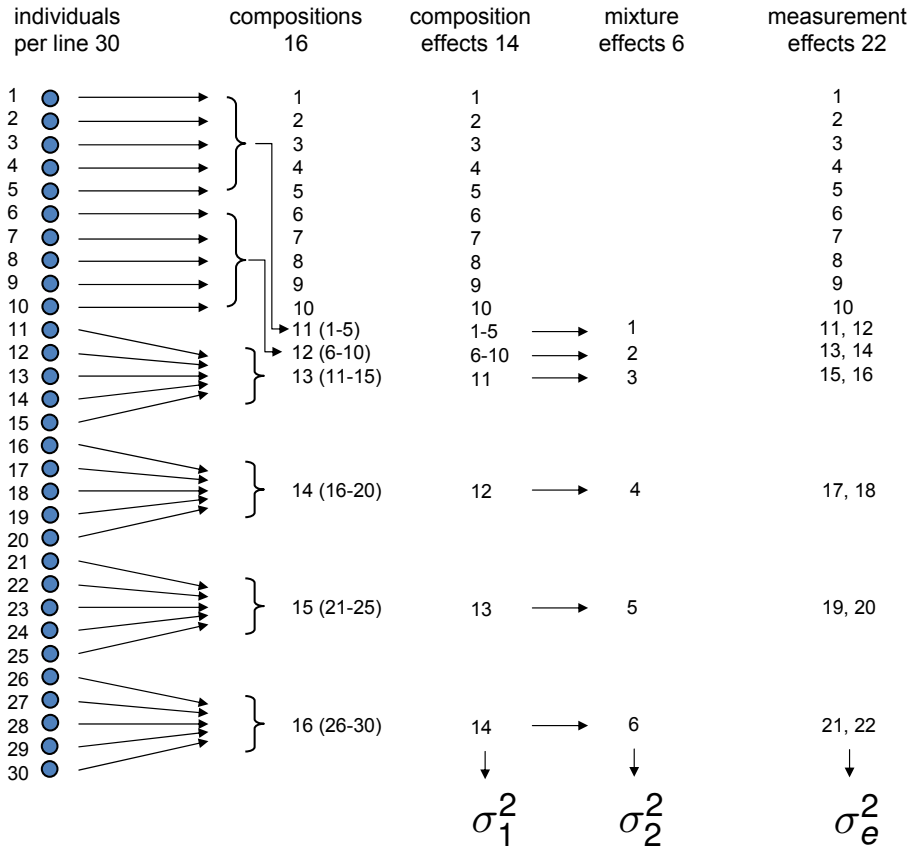


Figure 5.3: Random effects and variance components in the mouse experiment (details in section 3.2.2). For simplicity only one mouse line is shown, the structure of the other line was identical.

were measured as a single sample. They were subsequently blended into two pools of size 5 (composition 11 and 12). The other four pools were retrieved as a mixture of the individual samples no. 11–30. If generated according to the last example of section 2.5, the correlation matrix of effects with dimensions 16×16 would not be invertible, because of linearly dependent rows. This happens, because the variance of the sum of five individuals effects is five times the variance of their pool. Therefore, the number of composition effects (corresponding to biological variance) was

reduced to 14 by crossing out no. 11 and 12. Here, the rule applies, which states that the effects of pools that only consist of individuals (or smaller pools) that get an effect of its own, have to be canceled. In the corresponding rows of $\mathbf{Z}$ the weights of these smaller elements in the canceled effect appear (see matrices for mouse in section 7.1.2). In doing so, the matrix $\mathbf{G}_1$ became diagonal with entries of reciprocal pool size (individual samples are referred to as pool size one). With minor adaption for the matrix $\mathbf{Z}$, the structure $\mathbf{Z}_1\mathbf{G}_1\mathbf{Z}_1^\top\sigma_1^2$ was maintained and biological variance σ_1^2 could be estimated adequately. The first two pools (composition no. 11 and 12) gave an effect for the mixture of a pool, as well as the other four pools (compositions no. 13–16). Thus, six effects for blending of individuals were assigned and the corresponding blending error variance could be estimated, with respect to the correlation matrix $\mathbf{G}_2$ with the diagonal entries $\frac{\gamma-1}{\gamma^2} = \frac{4}{25}$. Since all pools were measured twice, in the end there were 22 arrays used per mouse line.

The second data set were from rats, which were analyzed in Kendziorski *et al.* (2005) separately for each pool size. For the third data set from honeybees, modeling had to be adapted because a two-color array platform was used. The vector of log-ratios $\mathbf{M} = \log\frac{R}{G} = \log R - \log G$ consists of the differences in intensities of the red and green channel. Therefore, the parameter vector $\boldsymbol{\beta} = (\mu, \Delta, b_{12}, b_{23})^\top$ also consists of differences for behavior (Δ) and tissue (b_{12}, b_{23}). The intercept μ includes the dye effect, i.e. the difference of the red and green channel. Random effects for each sample composition (u_1) and for the blending of individuals (u_2) are defined as before. Both design matrices for the random effects differ from the experiments with one-color arrays. Each row of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ now contains two non-zero elements in order to model the differences between effects, with entries of 1 for the red and -1 for the green channel. The same system was applied to the fourth data set of humans, who were analyzed for their susceptibility to tuberculosis with dependence of the tuberculosis skin test result.

As true proportions of aliquots are not available, and the relationship between deviations from homogeneous aliquots of individuals and the random effects of mixtures

of pools on the log-scale is complex, methodology was checked by simulation. The principle was to simulate normally distributed gene expressions of the individuals of a pool, transform it to the original scale, calculate mixtures as linear combinations with random weights, and transform back to the log-scale (see equation 3.5). For the simulation of weights the Dirichlet distribution was used with parameters $a_i = \frac{1}{\sigma_z^2} - \frac{1}{\gamma}$. To get weights with a certain variance σ_w^2 the pooling technical variance σ_z^2 can be calculated through the approximation $\sigma_z^2 \approx \frac{\gamma^3}{\gamma-1}\sigma_w^2$. It was shown in section 3.2.3, that the expectation of each weight was $\frac{1}{\gamma}$. Three different proportions of transcripts $(0, \frac{1}{3}, 1)$ were simulated, as affected by imperfect pooling. The first case (0) was to check the distribution of the REML log-likelihood ratio test (RLRT) statistic under the null hypothesis ($\sigma_2^2 = 0$). With the assumed half-half mixture of chi-squared distributions with zero and one degree of freedom, the distribution of p-values of the test statistic in figure 3.2 was uniform on the interval $[0, 0.5)$, as expected. Furthermore, simulations were executed to assess the impact of imperfect blending on differential expression. A design identical to the mouse experiment was chosen, with those estimated variance components. Differentially expressed transcripts were simulated and the performance in detection was investigated with the full model (m2) and the model (m1) by means of the type I and type II errors. Fit of models was done by EM-REML, and statistical analyses were performed for $\sigma_2^2 = 0$ with RLRT. False discovery rates of 5% for identification of certain transcripts with a non-zero blending error variance were applied. As a measure for relevance, the proportion of false null hypotheses was estimated by $\hat{\pi}_1 = 1 - \hat{\pi}_0$, where π_1 was taken from the R-package qvalue (Dabney *et al.*, 2011), and then the ability of both models to detect differential expression was compared with F-tests. The latter was determined from simulation results shown in figure 3.3. It turned out, that the level of permitted false discoveries 5% was exceeded by the model m1, but ensured in the full model m2.

The practical relevance of the second random effect for blending individual samples was investigated in experimental data of four different species, in chapter 3.

Methodology for estimation and tests of the corresponding blending error variance was developed in sections 3.2.1 and 3.2.4. With little modifications the model can be applied for analyses of data from both platforms, one- and two-color arrays. The hypothesis $\sigma_2^2 = 0$ was tested by RLRT, whose RLRT-statistic was calculated by taking twice the difference of the REML log likelihood between m2 and m1. The number of transcripts with a significant variance component due to imperfect blending was found to be 4329 (23%) in mouse data and 7093 (49%) in honeybees, but only six in rats and none whatsoever in human data. These results correspond to a false discovery rate of 5% in each data set. The relevance was further assessed by means of the estimated proportion of true null hypotheses. The proportion of transcripts, which were influenced by blending error variance, was determined as 0.75 in the mouse data set, 0.68 for bee, 0.29 for rat, and 0.40 for human. Differential expression was further compared for both models. The number of transcripts found to be differentially expressed between treatments was always higher when the blending error variance was neglected. Simulations clearly indicated overly-optimistic (anti-conservative) test results in terms of false discovery rates whenever this source of variability was not represented in the model (results shown in table 3.3). Therefore it can be concluded, that imperfect equality of shares, when blending RNA from different individuals into joint pools of variable size, is a source of technical variation with relevance for experimental design, practice in the lab and data analysis. Its potentially adverse effects, incorrect identification of differentially expressed transcripts and overly-optimistic significance tests, can be fully avoided, however, by the sound application of recently established theory and models for data analysis.

Modeling gene expression data adequately, especially from designs including pooling, in the opinion of the author has to deal with almost all sources of variation if possible. The aim was reached by modeling log expression differences with a mixed model, where two random effects corresponding to biological variance and blending error variance are assigned. The technical error (σ_t^2) is included in the residual variance σ_e^2. The possibilities with the proposed model are displayed in the figure

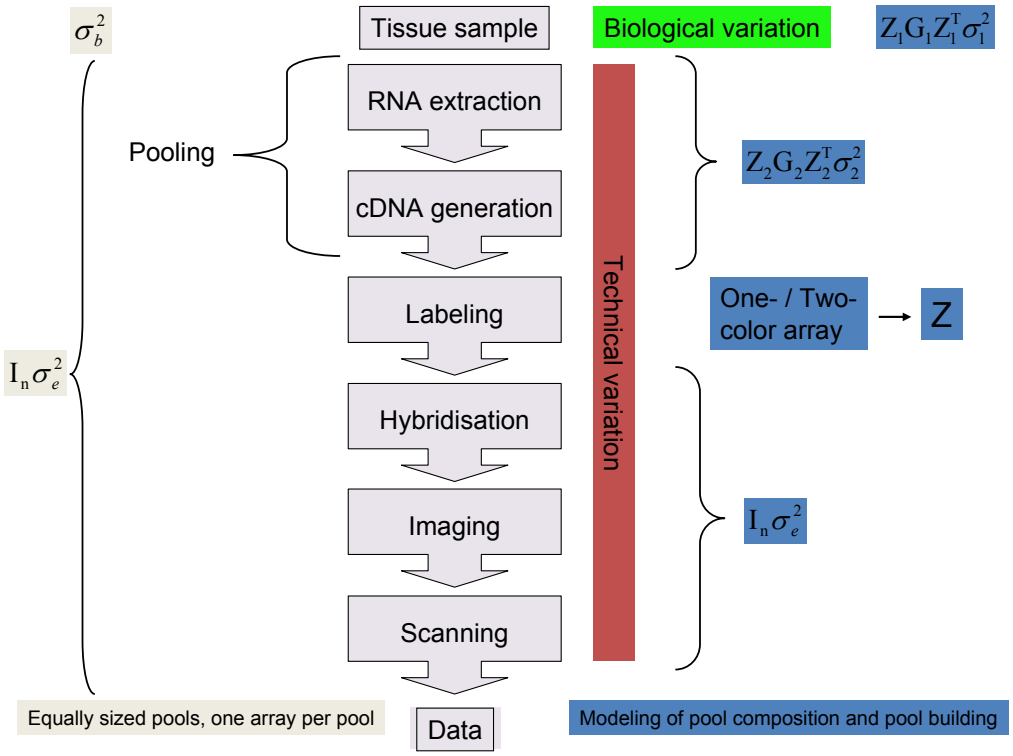


Figure 5.4: Scheme for steps of microarray experiment with pooling and sources of variation. On the left side, variance structure in experiments with uniform pool size. On the right side, the variance structure modeled in case of variable pool sizes with biological variance and blending error variance, with the option to account for one- or two-color platforms.

5.4 on the right side. Compared to a design with equally sized pools (left), the coverage of sources of variation is more detailed. Another advantage is that dye and measurement distortions — array by array as they appear — are associated with the residual errors.

5.3 Biomarker for hygienic behavior

Honeybee colonies are threatened by the mite *V. destructor*. Chemical treatment has been used following precisely developed plans (e.g. Rosenkranz *et al.*, 2010). There is consensus that breeding of tolerable colonies is more sustainable and more efficient than drug treatment. To enable the breeding of tolerable colonies, evaluation of hygienic performance of a colony of bees is necessary. Colony level measures — which are recommended for repeated capture — were developed, but require sev-

eral steps. These include the preparation of a comb section with damaged brood, and the determination of removal rate after a certain period of time. Identification of the individual hygienic behavior of bees, considered as uncapping and removal of diseased brood, is even more laborious. Therein, recorded video observations of individually marked bees have to be watched, and a detailed documentation of all tasks of bees on the particular section of the brood comb is required. A molecular marker based on behavior of individuals might be applied for marker-assisted selection. Furthermore, only some of the bees need to be collected for genetic analyses, which is user-friendly for beekeepers.

The development of a biomarker, considered as a set of genes with satisfactory discriminatory power whose joint expression pattern is predictive of class membership (Dziuda, 2010), is a goal in the investigation of hygienic behavior in the transkriptome. After using the experimental data from a preliminary experiment of HyBee in chapter 3, here the data from the main behavioral assay from the last project year was used, which also used two-color arrays, but with single samples. In order to select hygienic bee individuals, a section of a brood comb, artificially inserted with *Varroa* mites was presented to seven colonies. The activities of all the bees in this particular section were recorded in a 48 hours duration. The trait hygienic behavior was defined as a bee beginning uncapping ('beginner') and at least helping to remove ('helper') the diseased brood, exclusively. At the end of seven repeats (runs), 43 bees were observed showing this hygienic behavior. The intra-class correlation of the defined hygienic behavior among other traits was the highest, which means that a large part of variability in behavior can be claimed to the genotype (colony). An overview of the results of all the traits is given in the section 7.4. The 43 bees from four different colonies were compared with 52 controls in a dye balance design. In each group of bees (defined by colony and run), the hygienic bees had to be labeled equally with red and green dye. In groups with an odd number of hygienic bees, one of them was used a second time, but together with a new control. Furthermore, comparisons of hygienic bees with different genotypes were done with

12 arrays. Power of tests for differential expression was expected to benefit from this 'independent swap' over a standard dye swap, which was shown theoretically with variance functions in section 7.3.

Preparing data for biomarker search means to set a gene expression matrix (GEM) with one data point per sample. In our case, two-color array data (64 arrays) with p transcripts had to be transformed to sample (n=95) data. The GEM for two classes contains entries x_{ij}, which is the value for transcript $j\,(j = 1, \ldots, p)$ and sample $i\,(i = 1, \ldots, n_1 + n_2)$, where $n_1 = 43$ and $n_2 = 52$.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \ldots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_1+n_2 1} & x_{n_1+n_2 2} & \ldots & x_{n_1+n_2 p} \end{bmatrix}, \qquad \mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}^\top$$

Then, the data matrix consists of sample data points $\mathbf{x}_i$ (rows) of length p. If the rows are not sorted by classes, the class membership for a biomarker search can also be given as a dummy-coded vector.

The transformation of two-color array data (differences) with replications (multiple use of bees, correlations) is non-standard. Thereby, for an analysis with multivariate methods a good preservation of the variance structure is essential. A mixed model approach was developed, which allowed us to summarize all available information for each sample by the estimation of a random bee effect, which was then entered as a phenotypic value in the data matrix. The vector of differential log expression values $\mathbf{M}$ was modeled with a single fixed effect μ (the mean effect), which contains distortions through color. The design matrix $\mathbf{Z}$ is similar to the matrices for the composition of samples from chapter 3. The gene expression matrix is built column by column with the fitted random effects for bees. Of course, so far no class information has been used. To choose a transcript out of several thousand, the adaptive lasso regression was applied. Via cross-validation the optimal shrinkage parameter

and therewith the transcripts are chosen. In the same step, regression coefficients are determined. The search method is based on re-sampling. First the 64 arrays were split into 52 for the biomarker search, and 12 which were separated for an internal validation. Then, the preprocessing with the mixed model was executed separately for learning and validation data. The 73 bees from the first 52 arrays built the rows of a gene expression matrix for learning. Then, in 5000 repetitions they were sampled by 1/3 for a test set (24 bees), and 2/3 for a training set (49). The latter was used for the adaptive lasso fit, the selected features were stored for every repetition, and with the estimated coefficients the predictions of each test set of 24 bees were calculated. The classification of each test set was evaluated via the proportion of wrongly classified samples (prediction error). In the 5000 repetitions, the median was $\frac{5}{24}$. With the records of used transcripts frequencies of usage were calculated. Figure 4.4, which shows the corresponding histogram was used to determine the size of the biomarker candidate. Since the usage of transcripts was only slightly different beginning from rank 5, the first four transcripts were chosen. The coefficients for predictions were determined in a logistic regression with these four transcripts . Therewith, for the internal validation the preprocessed gene expression matrix from the 22 separated bees was used to calculate predictions, which are shown in figure 4.5. At the standard cutoff of 0.5, they were classified to the hygienic class or the non-hygienic class. The prediction error was found to be 27%. This is inside the confidence interval for four used features in figure 4.3. Thus, the average prediction error in the learning correctly assessed the prediction error in the internal validation set. Based on homology searches, the transcript most often used (AM03409), is involved in neuron development. This agrees with results of Navajas *et al.* (2008), where the set of genes differentially expressed between tolerant and sensitive bees were mainly involved in transcription and neuron development.

Different approaches for the selection of a biomarker candidate are of course possible. The selected biomarker candidate might be to some extent dependent on the method adaptive lasso. Correlation-based filter approaches or discriminatory mea-

sures like $\mathbf{T}^2$ (Dziuda, 2010) could be used instead. Also the principle to select from a top list according to the number of appearences in all 5000 resamplings might be changed. For instance, in Pinsky and Zhu (2011), it was shown that a highly negative correlated feature to the first feature is advantageous for performance. Such tuning methods are somewhat too advanced for the discovery phase of a biomarker search. A next step would be to compare different approaches and to check if they lead to similar results. A validation study should prove the predictive ability of a biomarker, with an independent data set of new samples from the target population. The application of a validated biomarker for beekeeping is possible, one could just collect bees and analyze them for their hygienic potential for a marker-based selection, but still difficulties could arise. Because of the low proportion of hygienic bees in a colony, a collection of say 50 bees per colony could be analyzed. If using pooled samples, costs would be controlled but methods have to be adapted. For predictions using pooled samples, a biomoarker search with pooled samples is also recommended. This biomarker is then not applicable for the prediction of single samples (Telaar *et al.*, 2013). Therefore, it is necessary to decide which biomarker, for pooled or single samples, is more useful for beekeeping.

The preprocessing approach presented allows access to much more data for a biomarker search, namely from two-color array experiments. As the principle to derive a phenotypic value for each individual is general, it can also be used with data from other platforms. For example, with RNA-sequencing data, but with respect to the Poisson distribution of count data, a generalized linear mixed model would be appropriate.

5.4 Using R for programming of analysis tasks

Difficulties arise if one wants to apply the full model with standard software packages like SAS or SPSS. The input of custom effects and variance components is difficult, and application of models to thousands of genes needs further (macro) programming. For estimation of mixed models the package ASReml (Gilmour *et al.*, 1995)

offers more functionality. It uses the average information REML algorithm, which improves the convergence while fitting the model in comparison to EM-REML. Correlation matrices can be specified, but need to be positive definite for algorithms. The design matrices $(\mathbf{X}, \mathbf{Z})$ are computed from the data. In doing so, it might happen that the inverse of the correlation matrix of random effects $(\mathbf{G})$ does not exist. In the end for each variance component j, a structure $\mathbf{Z}_j\mathbf{G}_j\mathbf{Z}_j^\top\sigma_j^2$ has to be fitted to the data, this provides another option for handling the problem of the inversion of $\mathbf{G}_j$. Defining the effects from the smallest disjunct elements, $\mathbf{G}_j$ is invertible and correlations are mapped through $\mathbf{Z}_j$. An example how this works can be seen for the mouse data in chapter 3. Therefore the option to enter matrices directly is important for the usability of analysis software.

To have all programming under the same platform, the open source software R was chosen for preparations, fitting, and analyses of the data. Inside R, several packages for analysis of gene expression data exist. For linear models, 'Limma' (Smyth, 2005) provides possibilities to analyze gene expression data with linear models. Therein, mixed models were used for normalization and estimation of spot correlations in a Bayesian approach. In Smyth and Altman (2013), a model for 'Estimating the intra-spot correlation from the M-value and A-values' contains effects for samples. So far 'Limma' does not provide an option to enter random effects for individuals when modeling the log gene expressions. Therefore, the R-script 'emreml.R' was created, for estimation of variance components, solution of effects and testing of contrasts. The possibility to directly enter all the matrices $(\mathbf{X}, \mathbf{G}, \mathbf{Z})$ needed was important. The output of the estimations with the R-script 'emreml.R' are shown in figure 5.5. It can be seen that differences between classes ('delta'), which are important for the question of interest, might differ between the two models m1 and m2 (see discussion above). The code of the emreml.R function is given in the appendix. There exists no R-package, to the author's best knowledge, which comfortably allows the analyses recommended in this work. Thus, future research directions could include the further development of the R-script which could be incorporated in an R-package.

	m1 loglh	m1 delta	m1 sigma_e^2	m1 sigma_1^2	m2 loglh	m2 delta	m2 sigma_e^2	m2 sigma_1^2	m2 sigma_2^2
1	42.09905	0.08825813	0.010430625	0.0878477145	49.83809	0.27718939	0.007450469	0.017085987	0.342758819
2	62.42431	0.06974972	0.009812557	0.0093139973	64.17506	0.09321161	0.006951189	0.007025708	0.045500943
3	60.57780	0.04609593	0.008579208	0.0161788990	62.11884	0.01527470	0.006673767	0.010962079	0.045166913
4	40.83996	0.45496675	0.017620502	0.0593901652	40.84118	0.45431160	0.017332623	0.059699639	0.002709609
5	41.00656	0.28450898	0.029795715	0.0218624342	42.01564	0.27755716	0.019938398	0.026551168	0.094539178
6	36.77707	0.15923954	0.049284243	0.0006439431	38.58250	0.20437227	0.027500383	0.014796522	0.184495727
7	39.04216	0.20273630	0.018607138	0.0673993706	42.85163	0.22559359	0.006973913	0.064352825	0.167653784

Figure 5.5: Output from emreml functions of mouse data set summarized for some transcripts (rows). On the left side, the estimations for the reduced model (with prefix m1) are presented, and on the right side, the estimations for the full model (m2). Note that, two times the difference of 'loglh' yielded the RLRT-statistic.

The strength of mixed models was used in this thesis to account for several issues of gene expression experiments and variance structure of the data. Phenotypic values are handled with regard to the influence of individuals; naturally, the individual appears as a random effect, if behavior on the basis of the transkriptome is declared. Problems with analysis in statistical software packages were circumvented by programming scripts for EM-REML and F-tests in linear mixed models.

6 Zusammenfassung

Ziel dieser Arbeit war es Versuchspläne und Auswertungsmethoden zu entwickeln und einzusetzen für Genexpressionsexperimente zum Vergleich von Bienen mit unterschiedlichem Hygieneverhalten. Seit einiger Zeit ist der Bestand an Bienenvölkern durch die Milbe Varroa destructor bedroht. Die Milben, welche sich in der Brut vermehren und von der Homolymphe der Bienenlarve ernähren, breiten sich im Stock aus und schwächen die Bienenvölker sehr stark. Die Milbenpopulation ist ein Hauptfaktor für das Zusammenbrechen der Bienenvölker. Die Bruthygiene der Bienen ist ein komplexes Verhaltensmuster und kann als natürlicher Abwehrmechanismus zur Reduzierung der Milbenpopulation beitragen. Nur wenige Bienen sind in der Lage Varroa-parasitierte Brut zu erkennen. Das Hygieneverhalten gegenüber der Milbe besteht aus dem Öffnen infizierter Zellen und dem Ausräumen der befallenen Larven. Die Erforschung der genetischen Komponenten dieses Verhaltens kann einen wichtigen Beitrag zur Zucht von Varroa-toleranten Bienenvölkern leisten. Dazu wurden Bienen, die das gewünschte Verhalten zeigten, in aufwendigen Verhaltensessays identifiziert. In sieben Durchgängen wurden den Arbeitsbienen von sieben verschiedenen Bienenvölkern Abschnitte mit Varroa-parasitierter Brut präsentiert. Die Auswertung der jeweils 48-stündigen Infrarot-Videoaufnahmen lieferte unterschiedliche Anzahlen der hygienischen Bienen für die verschiedenen Völker. Zur Reduzierung von Störgrößen wurde geplant nur Bienen des gleichen Volks und des gleichen Durchgangs hinsichtlich des Verhaltensunterschieds miteinander zu vergleichen.
Aufgrund der begrenzten Zahl von Arrays und da die Zahl der hygienischen Bienen unbekannt ist, war weiterhin überlegt worden, Mischproben der Bienen je nach dem

in welcher Zahl sie auftreten zu erzeugen. Dies führt zu variablen Poolgrößen, wovon in der bisherigen Literatur abgeraten wurde. Der erste der Hauptgründe sind die Verzerrungen zwischen den Messungen, welche durch die Log-Transformation bei der Normalisierung der Daten entstehen. Zweitens erfordert die erzeugte Varianzinhomgenität besonderen Aufwand bei der Modellierung und Auswertung der Messungen. Der interessierende Unterschied im Verhalten kann mit Hilfe von Kontrastfunktionen der geschätzten Parameter im gemischten linearen Modell analysiert werden. Im ersten Kapitel wurde gezeigt, dass das Ausbalancieren der Poolgrößen (Anzahl der Individuen in einer Mischprobe) zwischen den Gruppen die verzerrungsfreie Schätzung von Kontrasten ermöglicht. Für den allgemeinen Fall wurde eine Bedingung für die Verzerrungsfreiheit der Kontraste abgeleitet. Darin gehen die Designmatrix ($\mathbf{X}$), eine angenommene Varianzstruktur ($\mathbf{V}$), sowie die Kontrastfunktion ($\mathbf{K}$) ein. Die Varianzheterogenität wurde in einem gemischten linearen Modell durch zwei zufällige Effekte berücksichtigt. Im Modell $\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}$, ist das erstens die zufällige Abweichung (u_1) durch die Kombination von Individuen welche einer biologischen Varianz unterliegen. Außerdem wurde die Unsicherheit durch ungleiche Anteile der Individuen im Pool berücksichtigt, welche beim Mischen der Proben entsteht (u_2). Die dazugehörigen Varianzkomponenten (σ_1^2 und σ_2^2) gehen in die modellierte Varianzstruktur ein.

Im zweiten Abschnitt wurde die Bedeutung der neuen Varianzkomponente (σ_2^2, engl.: blending error variance) anhand experimenteller Daten untersucht. Dazu wurden Datensätze analysiert, in welchen verschiedene Poolgrößen vorkamen. Diese stammten von Mäusen, einem Vorversuch im Bienenprojekt, Ratten und Menschen. Bisher konnten diese nur separat nach Poolgröße, oder für einen Teil der Individuen mit fester Poolgröße ausgewertet werden. Der Einfluss der neuen Komponente wurde durch den Vergleich der Eigenschaften des vollen Modells und eines reduzierten Models, das diese nicht enthielt, untersucht. Mit Hilfe von REML-Likelihoodquotienten-Statistiken wurde geprüft, ob die Komponente signifikant verschieden von Null ist. Nach Korrektur der entsprechenden p-Werte auf multiples

Testen konnten in drei der vier Datensätze Transkripte identifiziert werden. Bei den Daten der Mäuse und Bienen waren diese sehr zahlreich, wenige wurden bei den Ratten gefunden, und keins bei den Human-Daten. Der Anteil der durch die Mischungsvarianz beeinflussten Transkripte wurde mit Hilfe des Anteils der wahren Nullhypothesen geschätzt. Außerdem wurden die differentiell expremierten Transkripte hinsichtlich des jeweiligen Gruppenvergleichs in den Experimenten mit beiden Modellen bestimmt. Hier zeigten sich Unterschiede, im reduzierten Modell wurden zumeist mehr Transkripte identifiziert, was durch einen höheren Fehler 1. Art hervorgerufen werden kann. In einer Simulationsstudie analog zum Design im Mäuse-Datensatz konnte gezeigt werden, dass das α-Niveau beim Testen im reduzierten Modell nicht eingehalten wird, aber im vorgeschlagenen Modell sind die Tests konservativ.

Die Genexpressionsdaten aus dem Hauptversuch (mit Einzelproben) des Bienenprojekts wurden im dritten Abschnitt dieser Arbeit für eine Biomarkersuche verwendet. Das Ziel hierbei war, ein Genexpressionsmuster mehrerer Transkripte zu entdecken, welches die Klassifizierung unbekannter Bienen in die Klassen hygienisch und nichthygienisch ermöglicht. Das ausgewählte Lernverfahren adaptive Lasso benötigt (wie die meisten anderen) eine Datenmatrix welche einen Datenpunkt je Individuum enthält. Daher ist eine Aufbereitung der in Differenzen der Log-Intensitäten von rot und grün gelabelten Proben vorliegenden Genexpressionsdaten nötig. Dazu wurde ein spezielles Vorbereitungsverfahren, wiederum mit Hilfe gemischter linearer Modelle, entwickelt. Dies ermöglicht es, die vorliegenden Informationen zu einem Individuum als phenotypischen Wert zusammen zu fassen. Untersuchungen der Korrelationen vor und nach dieser Transformation zeigten, dass die Varianzstruktur viel besser erhalten bleibt, als bei einem einfachen Zurückrechnen auf Werte der rot oder grün gelabelten Individuen. Nach Unterteilung der Bienen in den Validierungsdatensatz (22 Bienen) und die Lernstichprobe (73 Bienen) wurde der neue Ansatz zum Aufstellen der Genexpressionsmatrix (GEM) separat angewendet, durch Schätzen der Bieneneffekte für jedes der 13439 Transkripte. Aus der Lernstichprobe

der 73 Bienen wurde zufällig das Trainingssets gezogen und adaptive Lasso auf den entsprechenden Teil der GEM angewendet. Dabei werden Transkripte ausgewählt sowie deren Effekte geschätzt und benutzt um das Testset vorherzusagen. Die so geschätzte Klassenzugehörigkeit wird mit der tatsächlichen des Testsets verglichen und der Vorhersagefehler als Anteil der falschen unter allen Vorhersagen errechnet. Dieses Schema wurde 5000-fach wiederholt und dann die gespeicherten Transkripte anhand ihrer Häufigkeiten sortiert. Da nach den ersten 4 ein deutlicher Abstand in der Häufigkeit der selektierten Transkripte vorlag, wurde der Biomarker Kandidat hier auf 4 begrenzt. Mit diesen 4 Transkripten wurde ein einzelner Lernschritt mit der gesamten Lernstichprobe durchgeführt. Die Koeffizienten wurden anschließend, zur Vorhersage der unabhängigen 22 Bienen genutzt. Dabei ergab sich ein Vorhersagefehler von 27%, eine Spezifität von 83% und eine Sensitivität von 60%. Ein molekularer Marker kann eine wertvolle Hilfe bei der Zucht auf hygienische Bienen sein. Ein Vorteil wäre, dass der Imker nur ein paar Bienen zur Analyse einschickt, anstatt wie jetzt aufwendig und in regelmäßigen Abständen die Ausräumrate bei künstlich geschädigter Brut zu überprüfen. Dazu sind aber weitere Vorbereitungsschritte nötig, wie z.B. eine Validierungsstudie, oder je nach Einsatzgebiet (Analyse der Königin, Vorhersage mit gepoolten Proben) auch ein komplett neues Experiment und anschließende Lernverfahren.

Genexpressionsexperimente ermöglichen die Analyse komplexer Merkmale auf Ebene des Transkriptoms. Gegenüber SNP- oder Sequenzing-Techniken liegt ein Vorteil im Einblick in die gewebespezifischen Zusammenhänge, z.B. zeitnah wenn ein Individuum ein Verhalten ausführt. Obwohl es inzwischen neuere und teils genauere Verfahren zur Bestimmung des Transkriptoms gibt, sind Genexpressionsmessungen weiterhin etabliert, um einen Überblick zu liefern als Basisuntersuchung oder auch bei der Biomarkersuche, da ein Herausfiltern univariat unbedeutender Transkripte nachteilig für multivariate Verfahren ist. Die phenotypischen Werte auf Basis des Transkriptoms wurden modelliert, und der zufällige Einfluss des Individuums auf das Verhalten natürlicherweise als zufälliger Effekt im gemischten linearen Modell.

7 Appendix

7.1 Supplement for Chapter 3

7.1.1 Simulated and estimated blending error variance

In various simulation runs, the pooling technical variance σ_z^2 was altered in the range of $(0, 2.7]$ to evaluate whether the approximation in equation (3.3) is applicable for our purposes. Numbers of individuals in a pool were randomly from a Poisson

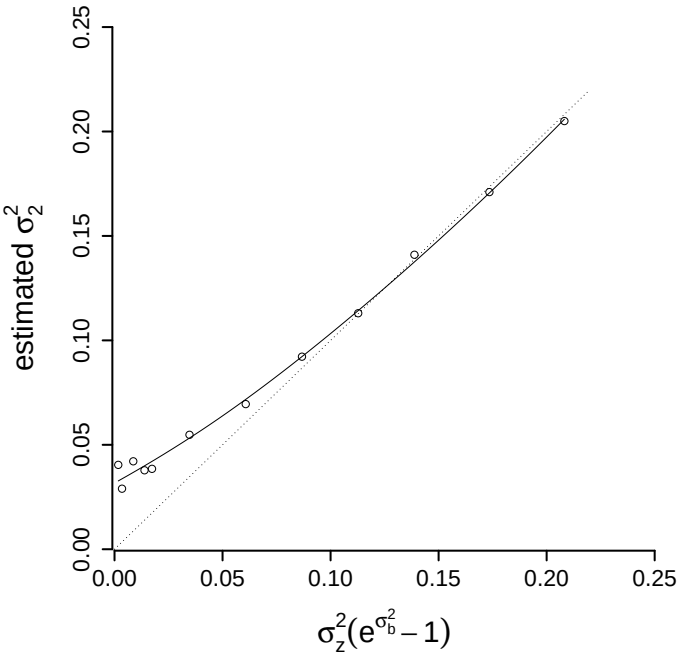


Figure 7.1: Plot of the average estimated variance components $\hat{\sigma}_2^2$ versus simulated values $\sigma_z^2(e^{\sigma_b^2} - 1)$.

distribution chosen. For each number, as many individuals were artificially blended

into a pool and an equally sized pool of controls was opposed. In figure 7.1 estimates and simulated values agree very well; some bias for small values can be attributed to the EM-REML algorithm used for variance component estimation.

7.1.2 Matrices for EM-REML and mixed model equations

This section shows various matrices for the experimental data sets in detail. They were explained to some extent in the section 3.2.2.

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

Figure 7.2: Design matrix $\mathbf{Z}_1^m$ for random composition effect of the mouse data.

$$
\begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&1
\end{bmatrix}
$$

Figure 7.3: Design matrix $\mathbf{Z}_2^m$ for random blending effect of the mouse data.

A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A14	B2	B3	B5	B6	B8	B9	B10	B11	B12	B13	B14	B15
0	0	1/3	0	0	0	0	0	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1/2	1/2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1/3	0	0	0	0	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1/2	1/2	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1/3	1/3	0	1/3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1/2	1/2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1/2	0	0	1/2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1/2	1/2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1/3	0	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1/2	1/2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	0	0	0	0	0	0	0	0	0	0	0	0
1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	0	0	0	0	0	0	0	0	0	0	0	0
1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	0	0	0	0	0	0	0	0	0	0	0	0
1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/3	1/3	0	1/3	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/2	1/2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/3	0	1/3	1/3
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/2	0	1/2	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/2	0	1/2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1/2	1/2	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/2	1/2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/2	1/2	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12
0	0	0	0	0	0	0	0	0	0	0	0	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12
0	0	0	0	0	0	0	0	0	0	0	0	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12
0	0	0	0	0	0	0	0	0	0	0	0	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12

Figure 7.4: Design matrix $\mathbf{Z}_1^r$ for random composition effect for the rat data. This data set was analyzed seperately by pool sizes in Kendziorski *et al.* (2005). On the top are displayed the original notations of the rats.

$$\mathbf{G}_1^h = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}, \mathbf{G}_2^h = \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{3}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{16} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{16} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{16} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{16} \end{bmatrix},$$

$$\mathbf{X}^h = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & -1 & -1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \mathbf{Z}_1^h = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{Z}_2^h = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 7.5: Design matrices of the honeybee data from the premilinary gene expression experiment of HyBee.

$$Z_2^g = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \end{bmatrix}$$

Figure 7.6: Design matrix $\mathbf{Z}_2^g$ for random blending effect of the human data from the GC6 project.

$$
G_1^g = \begin{bmatrix}
1/3 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 1/9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/9 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3
\end{bmatrix}
$$

$$
Z_1^g = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Figure 7.7: Correlation matrix $\mathbf{G}_1^g$ of biological effects and design matrix $\mathbf{Z}_1^m$ for random composition effect of the human data from the GC6 project.

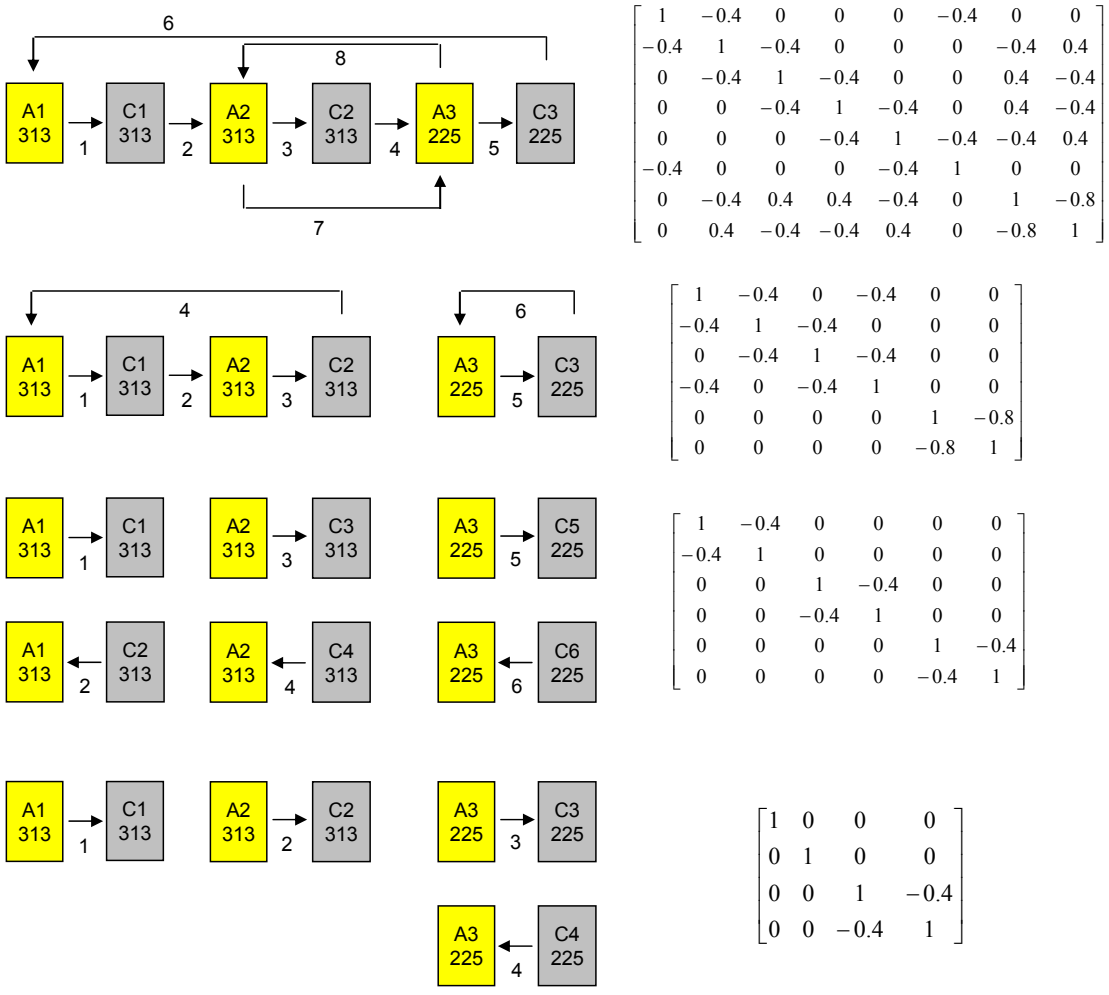
7.2 R-script for EM-REML

```
emreml=function (y, X, Z1, Z2, G1, G2, initE, initU1, initU2, display, cr){
  n <- length(y)
  Xy <- t(X) %*% y
  Z1y <- t(Z1) %*% y
  Z2y <- t(Z2) %*% y
  XX <- t(X) %*% X
  XZ1 <- t(X) %*% Z1
  XZ2 <- t(X) %*% Z2
  Z1X <- t(Z1) %*% X
  Z2X <- t(Z2) %*% X
  Z1Z1 <- t(Z1) %*% Z1
  Z1Z2 <- t(Z1) %*% Z2
  Z2Z1 <- t(Z2) %*% Z1
  Z2Z2 <- t(Z2) %*% Z2
  RHS <- c(Xy, Z1y, Z2y) ; lhsRow <- length(RHS)
  oldE <- initE ; oldU1 <- initU1 ; oldU2 <- initU2
  rankX <- qr(X)$rank
  G1inv <- solve(G1) ; G2inv <- solve(G2)
  rankG1 <- dim(G1)[1]
  rankG2 <- dim(G2)[1]
  z1 <- length(RHS)-dim(Z1Z1)[1]-dim(Z2Z2)[1]+1
  z2 <- length(RHS)-dim(Z2Z2)[1]+1
  diff <- 1 ; i <- 0
  while (diff > cr) {
    i <- i + 1
    lambda1 <- as.vector((oldE/oldU1))
    lambda2 <- as.vector((oldE/oldU2))
    LHS <- rbind(cbind(XX,XZ1,XZ2),cbind(Z1X,Z1Z1+G1inv*lambda1,Z1Z2),
                 cbind(Z2X,Z2Z1,Z2Z2+G2inv*lambda2))
    LHS_I <- solve(LHS)
    B <- LHS_I %*% RHS
    e <- y - cbind(X, Z1, Z2) %*% B
    sig2E <- (t(e) %*% y)/(n - rankX)
    c22 <- LHS_I[z1:(z2-1), z1:(z2-1)] ; c33 <- LHS_I[z2:lhsRow, z2:lhsRow]
    u1 <- B[z1:(z2 - 1)] ; u2 <- B[z2:length(B)]
    sig2U1 <- (t(u1) %*% G1inv %*% u1 + sum(G1inv * c22) * oldE)/(rankG1)
    sig2U2 <- (t(u2) %*% G2inv %*% u2 + sum(G2inv * c33) * oldE)/(rankG2)
    bn <- matrix(c(oldE, oldU1, oldU2), nrow = 3, ncol = 1)
    bn1 <- matrix(c(sig2E, sig2U1, sig2U2), nrow = 3, ncol = 1)
    diff <- sqrt(t(bn1 - bn) %*% (bn1 - bn))/sqrt(t(bn) %*% bn)
    if (display==TRUE) cat("iteration", i,"sol",format(B[1:(z1-1)],digits=4),
            "sige2",sig2E,"sigU12",sig2U1,"sigU22",sig2U2, "diff",diff,"\n")
    oldE <- as.numeric(sig2E)
    oldU1 <- as.numeric(sig2U1)
    oldU2 <- as.numeric(sig2U2)
  }
  V <- (Z1%*%G1%*%t(Z1))*as.numeric(sig2U1)+(Z2%*%G2%*%t(Z2))*
                         as.numeric(sig2U2)+diag(rep(1,n))*as.numeric(sig2E)
  Z <- cbind(Z1, Z2)
  V_ <- solve(V)
  LogLik <- -1/2*(log(det(V))+log(det(t(X)%*%V_%*%X))+
                  t(y)%*%(V_-V_%*%X%*%solve(t(X)%*%V_%*%X)%*%t(X)%*%V_)%*%y)
  beta <- t(B)[1:rankX]
  return(list(`Iterations`=i,`LogLh`=LogLik,`Solutions`=beta,
              `sigma_e^2`=sig2E,`sigma_U1^2`=sig2U1,`sigma_U2^2`=sig2U2))
}
```

The file displayed here, is an R-script for fitting a linear mixed model with two random effects and three variance components, namely biological, blending error, and residual variance. Here, it was used with input for: Vector of log-intensities $\mathbf{y}$, design matrices for fixed and random effects $\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2$, the (invertible) covariance matrices $(\mathbf{G}_1, \mathbf{G}_2)$, initial values for the variance components, display (TRUE/FALSE) for output of interim results, and the convergence rate cr (10^{-8} was used in this work).

7.3 Variance functions for comparison of designs

This section includes calculations of variance functions (vf) for several designs, done in preparation of a gene expression experiment with honeybees. In principle a lower variance function leads to a theoretically higher power to detect differential expression, because the inverse of the variance function is the middle term of the F-statistic for testing if $\mathbf{K\beta} = 0$. With lower vf larger F-statistics can be expected, which lead



$$\begin{bmatrix} 1 & -0.4 & 0 & 0 & 0 & -0.4 & 0 & 0 \\ -0.4 & 1 & -0.4 & 0 & 0 & 0 & -0.4 & 0.4 \\ 0 & -0.4 & 1 & -0.4 & 0 & 0 & 0.4 & -0.4 \\ 0 & 0 & -0.4 & 1 & -0.4 & 0 & 0.4 & -0.4 \\ 0 & 0 & 0 & -0.4 & 1 & -0.4 & -0.4 & 0.4 \\ -0.4 & 0 & 0 & 0 & -0.4 & 1 & 0 & 0 \\ 0 & -0.4 & 0.4 & 0.4 & -0.4 & 0 & 1 & -0.8 \\ 0 & 0.4 & -0.4 & -0.4 & 0.4 & 0 & -0.8 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -0.4 & 0 & -0.4 & 0 & 0 \\ -0.4 & 1 & -0.4 & 0 & 0 & 0 \\ 0 & -0.4 & 1 & -0.4 & 0 & 0 \\ -0.4 & 0 & -0.4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -0.8 \\ 0 & 0 & 0 & 0 & -0.8 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -0.4 & 0 & 0 & 0 & 0 \\ -0.4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -0.4 & 0 & 0 \\ 0 & 0 & -0.4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -0.4 \\ 0 & 0 & 0 & 0 & -0.4 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -0.4 \\ 0 & 0 & -0.4 & 1 \end{bmatrix}$$

Figure 7.8: Four possibilities for designs of a two-color array experiment (left) for measurement of three hygienic bees from two colonies, and corresponding theoretical covariance structure (right) using single samples. The yellow boxes symbolize samples of hygienic bees (denoted $A1, A2, A3$), gray was used for controls ($C1, \ldots, C6$). The latter number in the box named the genotype. Arrows symbolize two-color arrays, head (tail) the red (green) labeled sample. The covariance matrices shown, were calculated for biological variance of $\sigma_b^2 = 0.4$ and residual variance of $\sigma_e^2 = 0.2$.

to more rejections of the null hypothesis.

In figure 7.8 are shown minimal examples of four designs strategies. Three rare cases (hygienic bees) were assumed. The first (top) design compared the 6 samples with a design containing a loop over the samples (arrays no. 1 to 6), plus two arrays for comparison of genotypes (arrays no. 7 and 8), to secure comparison of hygienic (HY) and control (NH) bees without bias due to color and without distortion through colony. With the assumed variances ($\sigma_b^2 = 0.4$, and $\sigma_e^2 = 0.2$) covariance matrices were calculated as $\mathbf{V} = \mathbf{ZGZ}^\top\sigma_b^2 + \mathbf{I}_n\sigma_e^2$. The design matrices $\mathbf{X}$ were set with parameters according to differences of behavior (HY-NH) and differences in genotype (313-225). In the first case the design matrix was $\mathbf{X} = \left[\begin{smallmatrix} -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & -1 & 0 & 1 & -1 & 1 \end{smallmatrix}\right]^\top$, where x_{11} = -1 stands for effect NH-HY (C1(313) labeled red, A1(313) labeled green) on array 1. Furthermore, x_{12} was 0, because there were no difference in genotype on array 1. Then the variance function vf = $\mathbf{K}^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^-\mathbf{K}$ was found to be vf = 0.3. In the second design, loops were done only inside genotype, it turned out that vf = 0.3 was the same like before, but two arrays saved. The third design uses direct comparisons, where a kind of dye swap called 'independent swap' (same hygienic samples and new control, hybridized with opposite colors) was applied. Therewith we got 9 bees, and an improvement for vf = 0.233 was found. The fourth design also suffices for dye balance, with the minimum number of arrays (4). Here vf = 0.292 was calculated, which is still below the first two designs. Although results depend on assumptions for variance components, a dye balance design with an extra control to level the two numbers of red and green labeled hygienic bees can be recommended. In fact, design no. 4 was sequentially applied to groups of hygienic bees (same run and genotype) of the behavioral assays from the HyBee-project.

7.4 Intra-class correlations of hygienic tasks

The behavioral assay of HyBee was analyzed for several traits. A variety of activities of worker bees was available, broken down by the two cell types infested and non-infested. Workers which engaged the opening of a cell were denoted as 'beginner' and those who helped removal of brood as 'helper'. With a generalized linear mixed

Trait	Cell type	No. of bees	SumA> 2	ICC (%)
B & H	infested	52	8	51.3
Beginner	infested	68	1	16.5
Helper	infested	285	1	26.7
SumA	both	12888	113	22.5
B & H	control	67	18	9.7
B or H	control	593	18	16.4
Helper	control	391	0	18.5
Beginner	control	135	0	11.9
Active	both	1290	113	22.2

Table 7.1: Overview of investigated traits of the behavioral assays of HyBee. Activities of worker bees were recorded on video and were checked and evaluated by two observers. Each trait was analyzed seperately, hence the column 'No. of bees' accounts for all bees, whose activities included the trait at least once. The numbers in the column denoted SumA> 2 are the worker bees which showed the trait and were active at least 3 times. 'Beginner' is a bee, which engaged uncapping of a cell, and 'Helper' a bee involved in removal of brood.

model (SAS, proc glimmix) the intra-class correlations (ICC) for several traits were calculated. The model included a fixed effect for the run $(1, \ldots, 7)$ of the experimental setting and a random effect for the workers queen bee. The trait with the highest ICC was the detection/uncapping and removal. The results of table 7.1 means, that 51% of the variability in this trait can be attributed to the genotype of the queen bee.

Bibliography

Bienefeld, K., Ehrhardt, K., and Reinhardt, F. (2008). Noticeable success in honey bee selection after the introduction of genetic evaluation using BLUP. *American bee journal*, **148**(8), 739–742.

Boecking, O. and Spivak, M. (1999). Behavioral defenses of honey bees against varroa jacobsoni oud. *Apidologie*, **30**(2-3), 141–158.

Büchler, R., Berg, S., and Le Conte, Y. (2010). Breeding for resistance to varroa destructor in europe. *Apidologie*, **41**(3), 393–408.

Dabney, A., Storey, J. D., and with assistance from Gregory R. Warnes (2011). *qvalue: Q-value estimation for false discovery rate control*. R package version 1.32.0.

Dziuda, D. M. (2010). *Data mining for genomics and proteomics: analysis of gene and protein expression data*, volume 1. Wiley.com.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, **32**(2), 407–499.

Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, **19 (14)**, 1817–1823.

Gempe, T., Stach, S., Bienefeld, K., and Beye, M. (2012). Mixing of honeybees with

different genotypes affects individual worker behavior and transcription of genes in the neuronal substrate. *PLOS ONE*, **7**(2). e31653.

Genersch, E., Von Der OHE, W., Kaatz, H., Schroeder, A., Otten, C., Büchler, R., Berg, S., Ritter, W., Mühlen, W., Gisder, S., *et al.* (2010). The german bee monitoring project: a long term study to understand periodically high winter losses of honey bee colonies. *Apidologie*, **41**(3), 332–352.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**(4), 1440–1450.

Harbo, J. R. and Harris, J. W. (2009). Responses to varroa by honey bees with different levels of varroa sensitive hygiene. *Journal of apicultural research*, **48**(3), 156–161.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*, volume 2. Springer.

Jacobsen, M., Repsilber, D., Gutschmidt, A., Neher, A., Feldmann, K., Mollenkopf, H. J., Ziegler, A., and Kaufmann, S. H. E. (2007). Candidate biomarkers for discrimination between infection and disease caused by mycobacterium tuberculosis. *J Mol Med (Berl)*, **85**(6), 613–621.

Kendziorski, C., Irizarry, R. A., Chen, K. S., Haag, J. D., and Gould, M. N. (2005). On the utility of pooling biological samples in microarray experiments. *Proc Natl Acad Sci U S A*, **102**(12), 4252–7.

Kendziorski, C. M., Zhang, Y., Lan, H., and Attie, A. D. (2003). The efficiency of pooling mRNA in microarray experiments. *Biostatistics*, **4**(3), 465–77.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**(3), 983–997.

Kerr, M. K. (2003). Design considerations for efficient and effective microarray studies. *Biometrics*, **59**(4), 822–828.

Kessler, W. (2007). *Multivariate Datenanalyse: für die Pharma, Bio-und Prozessanalytik*. John Wiley & Sons.

Knapen, D., Vergauwen, L., Laukens, K., and Blust, R. (2009). Best practices for hybridization design in two-colour microarray analysis. *Trends in Biotechnology*, **27**(7), 406–414.

Kraemer, N. and Schaefer, J. (2012). *parcor: Regularized estimation of partial correlation matrices*. R package version 0.2-3.

Landgrebe, J., Bretz, F., and Brunner, E. (2004). Efficient design and analysis of two colour factorial microarray experiments. *Computational Statistics and Data Analysis*, **50**(2), 499–517.

Maertzdorf, J., Repsilber, D., Parida, S., Stanley, K., Roberts, T., Black, G., Walzl, G., and Kaufmann, S. (2010). Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes and immunity*, **12**(1), 15–22.

Mondet, F., Alaux, C., Severac, D., Rohmer, M., Mercer, A. R., and Le Conte, Y. (2015). Antennae hold a key to varroa-sensitive hygiene behaviour in honey bees. *Scientific reports*, **5**.

Mrode, R. A. and Thompson, R. (2005). *Linear Models for the Prediction of Animal Breeding Values*. CABI Publishing, 2nd edition.

Navajas, M., Migeon, A., Alaux, C., Martin-Magniette, M.-L., Robinson, G., Evans, J., Cros-Arteil, S., Crauser, D., and Le Conte, Y. (2008). Differential gene expression of the honey bee apis mellifera associated with varroa destructor infection. *Bmc Genomics*, **9**(1), 301.

Nota, B., Verweij, R. A., Molenaar, D., Ylstra, B., van Straalen, N. M., and Roelofs,

D. (2010). Gene expression analysis reveals a gene set discriminatory to different metals in soil. *Toxicological sciences*, **115**(1), 34–40.

Oxley, P. R., Spivak, M., and Oldroyd, B. P. (2010). Six quantitative trait loci influence task thresholds for hygienic behaviour in honeybees (apis mellifera). *Molecular ecology*, **19**(7), 1452–1461.

Passos, V. L., Tan, F. E. S., Winkens, B., and Berger, M. P. F. (2009). Optimal designs for one- and two-color microarrays using mixed models: a comparative evaluation of their efficiencies. *J Comput Biol*, **16**(1), 67–83.

Peng, X., Wood, C. L., Blalock, E. M., Chen, K. C., Landfield, P. W., and Stromberg, A. J. (2003). Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics*, **4**(26), 1–9.

Pinsky, P. F. and Zhu, C. S. (2011). Building multi-marker algorithms for disease prediction - the role of correlations among markers. *Biomarker insights*, **6**, 83–93.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rademacher, E. and Harz, M. (2006). Oxalic acid for the control of varroosis in honey bee colonies-a review. *Apidologie*, **37**(1), 98–120.

Rinderer, T. E., De Guzman, L. I., Frake, A. M., Tarver, M. R., and Khongphinitbunjong, K. (2014). An evaluation of the associations of parameters related to the fall of varroa destructor (acari: Varroidae) from commercial honey bee (hymenoptera: Apidae) colonies as tools for selective breeding for mite resistance. *Journal of economic entomology*, **107**(2), 516–522.

Rosa, G. J. M., Steibel, J. P., and Tempelman, R. J. (2005). Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comperative and Functional Genomics*, **6**, 123–131.

Rosenkranz, P., Aumeier, P., and Ziegelmann, B. (2010). Biology and control of varroa destructor. *Journal of invertebrate pathology*, **103**, 96–119.

Rudolf, H., Pricop-Jeckstadt, M., and Reinsch, N. (2013). Flexible pooling in gene expression profiles: design and statistical modeling of experiments for unbiased contrasts. *Statistical Applications in Genetics and Molecular Biology*, **12**(1), 1–16.

Rudolf, H., Nuernberg, G., Koczan, D., Vanselow, J., Gempe, T., Beye, M., Leboulle, G., Bienefeld, K., and Reinsch, N. (2015). On the relevance of technical variation due to building pools in microarray experiments. *BMC genomics*, **16**(1), 1027.

Rudolf, H., Gempe, T., Leboulle, G., Beye, M., Bienefeld, K., and Reinsch, N. (2016). Are there biomarkers for hygienic behavior of individual apis mellifera workers? *(manuscript)*.

Schaeffer, L. (1986). Estimation of variances and covariances within the allowable parameter space. *Journal of Dairy Science*, **69**(1), 187–194.

Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, **52**(7), 3283–3299.

Schudoma, C., Steinfath, M., Sprenger, H., van Dongen, J., Hincha, D., Zuther, E., Geigenberger, P., Kopka, J., Köhl, K., and Walther, D. (2012). Conducting molecular biomarker discovery studies in plants. *Methods in molecular biology (Clifton, NJ)*, **918**, 127–150.

Searle, S. R. (1971). *Linear models*. Wiley.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Smyth, G. K. and Altman, N. S. (2013). Separate-channel analysis of two-channel microarrays: recovering inter-spot information. *BMC bioinformatics*, **14**(165), 1–15.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, **31**(4), 265–273.

Spivak, M. and Reuter, G. S. (2001). Varroa destructor infestation in untreated honey bee (hymenoptera: Apidae) colonies selected for hygienic behavior. *Journal of Economic Entomology*, **94**(2), 326–331.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 479–498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomwide studies. *PNAS*, **100(16)**, 9440–9445.

Swallow, W. H. and Monahan, J. F. (1984). Monte carlo comparison of anova, mivque, reml, and ml estimators of variance components. *Technometrics*, **26**(1), 47–57.

Telaar, A., Repsilber, D., and Nürnberg, G. (2013). Biomarker discovery: classification using pooled samples. *Computational Statistics*, **28**(1), 67–106.

Tempelman, R. J. (2005). Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. *Veterinary Immunology and Immunopathology*, **105**, 175–186.

Thakur, R. K., Bienefeld, K., and Keller, R. (1997). Varroa defence behavior in apis mellifera carnica. *American Bee Journal*, **137**, 143–148.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Tsuruda, J. M., Harris, J. W., Bourgeois, L., Danka, R. G., and Hunt, G. J. (2012). High-resolution linkage analyses to identify genes that influence varroa sensitive hygiene behavior in honey bees. *PLOS ONE*, **7**(11).

Vanselow, J., Nuernberg, G., Koczan, D., Langhammer, M., Thiesen, H.-J., and Reinsch, N. (2008). Expression profiling of a high-fertility mouse line by microarray analysis and qpcr. *BMC Genomics*, **9**(307).

Villa, J. D., Danka, R. G., and Harris, J. W. (2009). Simplified methods of evaluating colonies for levels of varroa sensitive hygiene (vsh). *Journal of apicultural research*, **48**(3), 162–167.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8**(6), 625–637.

Wu, Z., Irizarry, R., Gentleman, R., Martinez Murillo, F., and Spencer, F. (2004). A model based background adjustment for oligonucleotide expression. *Journal of the American Statistical Association*, **99**(468), 909–917.

Yang, X. (2003). *Optimal design of single factor cDNA microarray experiments and mixed models for gene expression data*. Ph.D. thesis, Virginia Polytechnic Institute and State University.

Young, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, **3**, 579–588.

Zhang, W., Carriquiry, A., Nettleton, D., and Dekkers, J. C. (2007). Pooling mRNA in microarray experiments and its effect on power. *Bioinformatics*, **23**(10), 1217–24.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertationsschrift selbständig und ohne fremde Hilfe erstellt habe, unter Verwendung keiner weiteren als der angebenen Quellen.
Die vorliegende Arbeit wurde in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt.

Dortmund, Juli 2015