Investigating possibilities to predict milk phenotypes in *Holstein Friesian* cows based on a more complex model of the genotype-phenotype map

Dissertation

zur

Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

Promotionsgebiet Bioinformatik

Fakultät für Informatik und Elektrotechnik

der Universität Rostock

vorgelegt von Nina Melzer, geboren am 02.03.1982 in Leipzig

wohnhaft in Wismar

Rostock, 17. Mai 2013

urn:nbn:de:gbv:28-diss2014-0195-4

Gutachter:	Prof. Dr. Olaf Wolkenhauer	Universität Rostock
	Dr. Dirk Repsilber	Leibniz-Institut für Nutztierbiologie
	Prof. Dr. Thomas Martinetz	Universität zu Lübeck

Tag der Verteidigung: 30.04.2014

Abstract

The presented work covers a broad spectrum of investigations, where methods and approaches from the fields of bioinformatics, biostatistics, animal breeding, genetics and systems biology were used regarding the question of improved genetic value prediction for cattle considering molecular data in addition to SNP-genotypes from cattle.
In modern breeding it is common since 2010 to record genetic information based on genome-wide SNP marker data and performance traits in each generation. These kinds of data are used to estimate the genetic effects of each SNP-genotype within the population (parental generation). The estimated genetic effects are used in combination with the recorded SNP-genotypes of the offspring generation to estimate their genetic values (breeding values), which represent the basis for the selection of the next parental generation. The estimation of genetic effects is based on a linear model, which can be termed the classical model of the genotype-phenotype map (GP map). Various estimation methods in the field of genomic selection are based on the classical model. To improve and to optimize such estimation methods, it is common to simulate data based on this classical model.
The main objective of this thesis was to investigate if an improved genetic value prediction can be obtained when an intermediate level of the GP map, in our case the metabolome, is considered. For this purpose, the different levels (genotype, metabolome, phenotype) were first simulated. In particular, these simulations were based on the data structure of the corresponding experimental data (from around 1,300 Holstein Friesian cows), which were collected. This allowed later a purely conceptual comparison between the simulated data and the experimental data regarding the observed prediction results. The experimental data collection based on different designs developed in this work and was supported by a specially created relational database. To realize the simulation of the metabolome a parameterized metabolic network for erythrocytes (SMBL model) was used. In this metabolic network some of the enzyme parameters were modeled as a function of varying SNP-genotypes, i.e., the simulated metabolite profiles depend on the SNP-genotypes. This systems biology approach enabled us to simulate genetic effects implicitly defined by the metabolic network. The phenotypes were simulated by adding up the obtained metabolic concentrations and adding a random error. The simulated three levels also allowed investigation of the expected degree of improvement of the genetic value prediction when the whole simulated metabolome or just a part of it is used. For this purpose an integrative bioinformatics approach was developed and implemented. This approach is divided into three steps: metabolome-phenotype, genotype-metabolome, genotype-phenotype. For the first step, methods from the field of statistical learning (e.g.,

random forest) were used, in order to enable variable selection, i.e., to obtain a measure of importance for each metabolite relating to the prediction of the phenotype. In the second step, the importance of each metabolite was used to weight corresponding SNP-genotypes, resulting in weighting of some genome regions. Finally, the weighted SNP-genotypes were used to predict the genetic value using a genomic selection method (Bayes approach, fast-BayesB). Results from the simulation study revealed that an improvement of the genetic value prediction is possible, especially if the whole simulated metabolome was used. The proposed integrative bioinformatics approach was also applied on the experimental data, wherein in the second step a Bayes approach (SVS) was used, which contains a variable selection method. Milk metabolites are selected which show high variable importances in the prediction of the milk trait. SNP-genotypes were selected which show a significant impact on these selected milk metabolites. The integrative bioinformatics approach resulted in a strongly reduced number of SNP-genotypes, which were used for the genetic value prediction (SVS). The respective results were compared with the classical approach, which revealed that comparable prediction precisions were obtained for the milk trait fat content for both approaches. Significance and relevance of selected SNP-genotypes using the new integrative bioinformatics approach were investigated in detail. In particular, milk metabolites and milk traits as well as the relationship between them were deeper investigated using univariate and multivariate analysis methods, wherein new associations between milk metabolites and milk traits were revealed. Considering the additional level of the GP map the metabolome allows further investigation of the relationship between the various levels, whose exploitation can lead to improved prediction of the genetic value.

The presented results in this thesis are of importance from a methodological and biostatistical point of view. In addition, they are of relevance from a zootechnical-biological perspective.

Zusammenfassung

Die vorliegende Arbeit umfasst ein breites Spektrum an Untersuchungen, wobei Methoden und Ansätze aus den Bereichen der Bioinformatik, Biostatistik, Tierzucht, Genetik und Systembiologie angewendet wurden mit dem Ziel die genetische Wertvorhersage beim Rind zu verbessern durch zusätzlich zu SNP-Genotypen mit einbezogene molekularbiologische Daten.

In der Rinderzucht ist es seit 2010 üblich, genetische Informationen in Form von genomweiten SNP Markerdaten und Leistungsmerkmale von jeder Generation aufzuzeichnen. Anhand dieser Daten können innerhalb einer Population (Elterngeneration) die genetischen Effektgrößen für jeden SNP-Genotyp geschätzt werden. Diese werden zusammen mit den ermittelten SNP-Genotypen der Nachkommen genutzt, um deren genetischen Wert (Zuchtwert) zu schätzen, der die Grundlage für die Selektion der nächsten Elterntiere bildet. Die Schätzung der genetischen Effektgrößen basiert auf einem linearen Modell welches als klassisches Modell der Genotyp-Phänotyp (GP) Abbildung bezeichnet werden kann. In dem Gebiet der genomischen Selektion existieren verschiedene Schätzmethoden, die auf dem klassischen Modell beruhen. Um Schätzmethoden zu optimieren und zu entwickeln werden Daten basierend auf diesem Modell simuliert.

Das zentrale Ziel der vorliegenden Arbeit ist zu untersuchen, ob eine verbesserte Vorhersage des genetischen Wertes erzielt werden kann, wenn eine weitere Zwischenebene der GP Abbildung, in diesem Fall das Metabolom, berücksichtigt wird.

Dazu wurden die Daten der verschiedenen Ebenen (Genotyp, Metabolom, Phänotyp) zunächst in Anlehnung an die Datenstruktur der entsprechenden experimentell erhobenen Daten (von rund 1.300 Holstein Friesian Kühen) simuliert, um später einen rein konzeptionellen Vergleich zwischen den Vorhersageergebnissen basierend auf den simulierten und experimentellen Daten zu ermöglichen. Die experimentelle Datenerhebung basiert auf verschiedenen in dieser Arbeit entwickelten Designs und wurde unterstützt durch eine eigens erstellte relationale Datenbank. Für die Simulierung des Metaboloms wurde ein parametrisiertes metabolisches Netzwerk für Erythrozyten (SBML Modell) verwendet, wobei einige Enzymeigenschaften als Funktion variierender SNP-Genotypen modelliert wurden, d.h. Metabolitprofile wurden in Abhängigkeit der SNP-Genotypen simuliert. Dieser systembiologische Ansatz ermöglicht es, die verschiedenen genetischen Effektgrößen implizit mittels des Netzwerkes zu simulieren. Die Phänotypen wurden simuliert, indem die entsprechenden Metabolitkonzentrationen aufsummiert und mit einem zufälligen Fehler versehen wurden. Die drei simulierten Ebenen erlaubten es zu untersuchen, inwieweit eine Verbesserung der Vorhersage des genetischen Wertes erzielt werden kann, wenn das gesamte oder nur ein Teil des simulierten Metaboloms berücksichtigt wird.

Hierzu wurde ein integrativ bioinformatischer Ansatz entwickelt, der in drei Schritte unterteilt ist: Metabolom-Phänotyp, Genotyp-Metabolom, Genotyp-Phänotyp. Im ersten Schritt wurden Methoden aus dem Gebiet des statistischen Lernens (z.B. Random Forest) angewendet, um per Variablenselektion ein Wichtigkeitmaß für jeden Metaboliten im Bezug auf die Vorhersage eines Phänotyps zu erhalten. Im zweitem Schritt wurden die Metabolit-Wichtigkeitsmaße zu den entsprechenden SNP-Genotypen zugeordnet, um bestimmte Genomregionen stärker zu gewichten. Die Vorhersage des genetischen Wertes erfolgte mit den gewichteten SNP-Genotypen unter Verwendung einer genomischen Selektionsmethode (Bayes Verfahren, fastBayesB). Die Ergebnisse zeigten, dass eine Verbesserung der Vorhersage des genetischen Wertes möglich ist, insbesondere wenn das gesamte simulierte Metabolom verwendet wird. Der integrativ bioinformatische Ansatz wurde auch auf die experimentellen Daten angewendet, wobei im zweiten Schritt ein Bayes Verfahren (SVS) angewandt wurde, das ebenfalls Variablenselektion ermöglichte. Dadurch konnten SNP-Genotypen selektiert werden, die einen signifikanten Einfluss auf bedeutsame Metaboliten für das untersuchte Milchmerkmal besitzen. Die stark reduzierte Anzahl an SNP-Genotypen wurde schließlich zur Vorhersage des genetischen Wertes benutzt und mit dem klassischen Auswertungsansatz verglichen. Hierbei zeigte sich, dass vergleichbare Präzisionen für das Milchmerkmal Fettgehalt für beide Ansätze erhalten wurden. Die mit dem neuem integrativ bioinformatischen Ansatz selektierten SNP-Genotypen wurden im Einzelnen auf ihre Signifikanz und Relevanz getestet. Insbesondere die Milchmetaboliten und die Milchmerkmale, sowie die Beziehung zwischen diesen beiden Ebenen wurden mittels univariater und multivariater Analysemethoden genauer untersucht, wobei neue Assoziationen zwischen Milchmetaboliten und Milchmerkmalen detektiert wurden. Die Berücksichtigung des Metaboloms im Modell der GP Abbildung ermöglicht weiterführende Untersuchungen der Beziehung zwischen ihren einzelnen Ebenen, deren Ausnutzung wiederum zu einer verbesserten Vorhersage des genetischen Wertes führen kann.

Die präsentierten Ergebnissen in dieser Arbeit sind dabei vorwiegend aus methodisch-biostatistischer Sicht von Bedeutung, aber auch aus tierzüchterisch-biologischer Sicht von Relevanz.

Contents

Abstract v

Zusammenfassung ix

List of Figures xvii

List of Tables xix

Acronyms xxi

Notation xxiii

1 Introduction 1

1.1 General objective . . . 1
1.2 Basics of dairy cattle breeding . . . 3
1.3 Background of cattle breeding . . . 9
1.4 Genomic selection methods . . . 14
1.5 Basic information on the bovine genome, QTL and cow's milk . . . 16
1.5.1 Bovine marker maps . . . 16
1.5.2 Quantitative trait loci in Holstein dairy cattle . . . 17
1.5.3 Cow's milk properties . . . 18
1.6 Bioinformatics and Systems Biology . . . 20
1.6.1 Omics . . . 20
1.6.2 Bioinformatics . . . 21
1.6.3 Systems Biology . . . 22
1.6.4 Metabolic networks . . . 23
1.7 Thesis objectives and structure . . . 25

2 Experimental data acquisition 29

2.1 Introduction . . . 29
2.2 Data collection . . . 30
2.2.1 From blood sample to genotype . . . 31
2.2.2 From milk sample to milk phenotype . . . 34
2.2.3 From milk sample to milk metabolite profile . . . 35
2.2.4 Results of experimental data preparation . . . 36
2.3 Randomization design for milk metabolite profiling . . . 37

2.4 BovIBI database . . . 39
2.5 Summary . . . 40

3 Analyses of simulated SNP-, metabolome-, and phenotype data 43
3.1 Introduction . . . 43
3.1.1 State of the art of data simulation . . . 44
3.1.2 Implementation of data simulation . . . 45
3.2 Material and Methods . . . 46
3.2.1 Simulation approaches to obtain a suitable LD . . . 46
3.2.2 The alternative (SBML) approach for simulation . . . 49
3.2.3 The metabolite approach for prediction . . . 55
3.3 Results for simulation studies . . . 57
3.3.1 Analysis of the simulation approaches regarding a suitable LD . . . 57
3.3.2 Conventional versus SBML simulation approaches . . . 57
3.3.3 Analysis of the metabolite approach for prediction . . . 63
3.4 Discussion . . . 64
3.4.1 Simulation approaches to obtain an appropriate LD . . . 65
3.4.2 Conventional approach versus SBML approach for simulation . . . 66
3.4.3 The benefit of using the metabolite approach for prediction . . . 68
3.5 Summary . . . 69

4 Analyses of experimental data 71
4.1 Introduction . . . 72
4.1.1 Background information for the analysis of the three system-levels 72
4.1.2 Analysis of the three system-levels . . . 74
4.2 Material and Methods . . . 75
4.2.1 Conceptual comparison between simulated and experimental data 75
4.2.2 Cross-validation designs . . . 77
4.2.3 Investigations of milk metabolites and milk traits . . . 78
4.2.4 The metabolite approach for experimental data . . . 83
4.3 Results of investigations of the experimental data set . . . 88
4.3.1 Analysis of experimental and simulated data . . . 88
4.3.2 Analysis of milk metabolites and milk traits . . . 89
4.3.3 Comparison of the metabolite approach to other approaches . . . 98
4.4 Discussion . . . 102
4.4.1 Experimental data versus simulated data . . . 104
4.4.2 Investigations of relations of milk metabolites and milk traits . . . 105
4.4.3 The metabolite approach compared to three other approaches . . . 109
4.4.4 Additional aspects of modeling . . . 113
4.5 Summary . . . 114

5 Conclusions 117

Bibliography 123

Appendix 141

A Additional information for fastBayesB 141
A.1 γ-values which were not suitable for fastBayesB 141
A.2 Goodness of model fit for simulated data 142

B Additional information about experimental data analyses 143
B.1 Correlation between milk metabolites and milk traits 143
B.2 Known QTL regions determined by use of the cattleQTL database 151
B.3 Trace plots for selected SNPs using SVS for three investigated milk traits. 156
B.4 Important milk metabolites detected using RF and PLS 157
B.5 Goodness of model fit for experimental data 161
B.6 Important SNP positions detected using the metabolite approach 162
B.7 Investigations of the importance of DGAT1 on three selected milk traits. . 164

Acknowledgement 167

Erklärung 171

Publications and own contributions 173

Theses 181

List of Figures

1.1 Scheme of a conventional progeny testing program. 11
1.2 Comparison of the duration of two breeding programs. 14
1.3 Factors influencing milk. 18
1.4 Idealized schematic representation of the milk yield during lactation. . . . 20

2.1 Overview of the experimental data preparation to obtain the desired three system-levels. 32
2.2 An example for each quality and quantity test for DNA. 33
2.3 The mass to charge ratios (m/z) of molecule fragments for the FAMEs presented over all measured milk samples. 35
2.4 An example metabolite spectrum (1,6-anhydro-beta-Glucose). 36
2.5 Overview of the number of collected milk samples. 38
2.6 Schematic representation of the milk randomization design. 41
2.7 Schematic representation of the BovIBI database. 42

3.1 Schematic representation of the conventional approach and the SBML approach. 52
3.2 Schematic representation of the weighting approach. 56
3.3 The mean LD values over 2,000 generations of ten replicates for the settings without mutation and for both scenarios with mutation. 57
3.4 The mean LD values over 2,000 generations of the ten replicates for mutation scenario 1 and mutation scenario 2 using different mutation rates. 58
3.5 Comparison of linkage disequilibria between experimental and simulated data. 59
3.6 The estimated main genetic effect sizes for the conventional approach and the SBML approach. 60
3.7 The estimated main genetic effect sizes for all QTL for each metabolic outcome. 61
3.8 The observed precisions of genetic value prediction are shown using different weights. 63

4.1 Scheme of the invariable double 10-fold cross-validation design. 78
4.2 Analysis design: schematic representation of the workflow. 85
4.3 Estimated main genetic effects for different milk traits. 89

4.4 An example of an increased trend and decreased trend over lactation days in milk metabolites. 90
4.5 The dendrogram resulting from hierarchical clustering of average metabolite profiles for the influencing factor lactation interval. 91
4.6 The dendrogram resulting from hierarchical clustering of average metabolite profiles for the influencing factor farm as well as the numbers of metabolites with significant differences between farms. 92
4.7 Correlations between milk metabolites and milk traits. 94
4.8 Correlations between milk traits. 95
4.9 Boxplots of observed precisions for the prediction of milk traits from metabolite profiles using RF and PLS. 96
4.10 Boxplots of the precision of the genetic value prediction of ten outer cross-validation runs for all tested approaches. 103

B.1 Trace plots for selected SNPs. 156
B.2 Residual plots for investigated milk traits using SVS. 161

List of Tables

2.1 Overview of the substance classification of the obtained metabolites. . . . 37
2.2 Part of the milk metabolite measuring design. 39

3.1 Used chromosome lengths for the simulation study for all 30 chromosomes of *Bos taurus*. 47
3.2 The enzyme characteristics which were changed in the SBML model. . . . 53
3.3 The average estimated variance components and corresponding simulated variance components are listed for both simulation approaches. 62

4.1 Estimated narrow-sense heritabilities for investigated milk traits. 82
4.2 Estimated variance components and prediction precisions for simulated data sets and chosen milk traits with fastBayesB. 88
4.3 The number of metabolites and milk traits, on which influencing factors impact significantly. 90
4.4 The number of metabolites with the most significant differences between levels of an investigated factor of interest. 93
4.5 Important metabolites detected for the factors lactation interval and farm. 95
4.6 For each milk trait, the observed important metabolites are listed in alphabetical order. 97
4.7 Information about important milk metabolites. 100
4.8 The average number of selected important SNPs. 101
4.9 The P-values from rating the important metabolites for the reduced classical approach and the metabolite approach. 102

A.1 The number of replicates leading to non-convergence and aborting rates over the 100 replicates for each tested scenario for both simulation approaches. 141
A.2 The mean correlation between fitted values and residuals is listed for all tested scenarios for both simulation approaches. 142

B.1 Pearson correlation matrix of milk metabolites and milk traits. 144
B.2 Known QTL regions filtered from the cattleQTL database. 152
B.3 All important metabolites using RF and PLS for each milk trait. 157
B.4 Detected important SNPs using the metabolite approach for investigated milk traits. 162
B.5 The mean prediction precisions are listed for all SNPs, without DGAT1-SNP and without DGAT1-region for all three investigated milk traits. . . 164

Acronyms

ANOVA	variance analysis
bp	basepair
BLUP	best linear unbiased prediction
cM	centiMorgan
DNA	desoxyribonucleic acid
EB	energy balance
FAME	fatty acid methyl ester
FBA	flux-balance analysis
FDR	false discovery rate
GC-MS	gas chromatography-mass spectrometry
GP map	genotype-phenotype map
GS	genomic selection
GWAS	genome wide association studies
HWE	Hardy-Weinberg equilibrium
LD	linkage disequilibrium
LDA	linear discriminant analysis
LE	linkage equilibrium
LRT	likelihood ratio test
M	Morgan
MAF	minor allele frequency
MAS	marker assisted selection
MCMC	Markov-Chain Monte-Carlo
MPT	milk performance test
ODE	ordinary differential equation
PLS	partial least squares regression
RF	random forest
SCC	somatic cell count
SCS	somatic cell score
SFA	saturated fatty acids
SNP	single nucleotide polymorphism
SVS	spike and slab variable selection
QTN	quantitative trait nucleotide
QTL	quantitative trait locus

Notation

General Note

In this thesis "ˆ" stands for the corresponding estimated components for the following listed symbols.

The following list contains symbols frequently used in this thesis.

List of Symbols

α	significance level
θ	recombination rate
γ	proportion of QTL to SNPs
h^2	narrow-sense heritability
H^2	broad-sense heritability
N	population size
n	number of animals
N_d	number of dams
N_{eff}	effective population size
N_s	number of sires
n_{SNP}	number of SNPs
n_{QTL}	number of QTL
m	mutation rate
w	weights
r^2	linkage disequilibrium
ρ	correlation between estimated and simulated genetic value
$\rho_{simulated}$	correlation between estimated genetic value and simulated phenotype
$\rho_{experimental}$	correlation between estimated genetic value and corrected observed milk trait
ρ_{milk}	correlation between predicted and observed milk trait
ρ_{SVS}^{milk}	correlation between estimated genetic value and observed milk trait using SVS
ρ_R	ρ_{SVS}^{milk} based on the resampling approach
ρ_{t1t2}	correlation between trait1 and trait2
σ_a^2	additive genetic variance
σ_d^2	dominance genetic variance
σ_e^2	residual variance
σ_g^2	genetic variance
σ_i^2	epistatic genetic variance
σ_p^2	phenotypic variance

σ_s^2	sire variance
V_{max}	enzyme kinetic maximum reaction velocity
y^{conv}	simulated phenotype based on conventional approach
y^{sbml}	simulated phenotype based on SBML approach

1 Introduction

1.1 General objective

This thesis investigates if additional information of the metabolome level can improve the genetic value prediction in dairy cattle. To address this issue, different methods and knowledge from several research fields were applied, e.g., bioinformatics and biostatistics. In the field of dairy cattle science, since the year 2010, it is common to record the genotypic and phenotypic information of an animal. Information on traditional milk traits (quantitative traits, e.g., milk fat, somatic cell count (SCC)) are important phenotypes collected routinely from cows. These milk traits are accessed via the standard milk performance test (MPT), which is carried out monthly for each dairy cow. The MPT is used to monitor the quantity and quality of milk. In this context, it is also of great interest to improve the detection and prevention of diseases (e.g., mastitis) and to monitor specific traits related to the state of health and management. The traditional milk traits used as biomarkers for the state of health are, however, not sufficiently sensitive in view of diagnostic efficiency.

The genetic information of an animal is assessed by using genome-wide marker data, which consist mostly of single nucleotide polymorphism (SNP) markers. Based on genotypic and phenotypic information, it is possible to estimate the genetic effects of the markers within a population (parental generation). The estimated genetic effects combined with the genetic information of the offspring generation are used to estimate the genetic values (considering additive and non-additive genetic effects) or breeding values (considering additive genetic effects). In general, the breeding value of an animal is estimated because it serves as a basis to decide whether the animal is used for breeding relating to a specific breeding goal. The use of genotypes and phenotypes for the estimation of genetic values is also referred to as the classical genotype-phenotype map (GP map), which excludes the consideration of further known intermediate levels (e.g., proteome, metabolome). To estimate the genetic effects within a population, various estimation methods exist in the field of genomic selection (GS). In this field, it is also common to use simulated data to compare different methods of genetic evaluation and to optimize methods, whereby a simple linear function is typically applied to the classical GP map of simulated data. In this thesis the approach based on the classical GP map for simulating data is termed conventional approach and the analysis based on the classical GP map is termed classical approach.

It is, however, not clear if the conventional approach is an appropriate basis to optimize such estimation methods in regard to experimental data. At the start of this thesis,

the question of whether including an intermediate level of the GP map, for example the milk metabolome, as an additional information source for the genetic value prediction might be beneficial had not yet been scientifically examined. Apart from the mentioned MPT, performed for many years (since the early 1950s), it is now possible to analyze milk metabolites in a high throughput manner and to identify functionally important metabolites, which can serve as biomarker candidates. In this context, only few publications can be found in the recent literature in which milk metabolites were investigated (discussed in Chapter 4). In these studies, mostly the correlation between single or groups of metabolites and single milk traits of interest were investigated. Out of these studies only few milk metabolites are proposed to be used as biomarkers. Analyses regarding multivariate correlations between sets of metabolites and milk traits from the MPT are still missing.

The main objective of this thesis was to investigate if an improvement can be achieved when the intermediate level of the GP map, in this thesis the metabolome is additionally considered for the genetic value prediction in Holstein Friesian cows. To allow investigations of the above mentioned issues, each system-level (genotype, metabolome, phenotype) was simulated on the one hand, and experimental data of these system-levels from about 1,300 Holstein Friesian cows were collected on the other hand. In total, 11 milk traits were measured in the standard MPT and 190 milk metabolites could be determined. The main topic of this thesis can be divided into three sub-topics, in which different aspects are investigated.

- The first task was to investigate the prediction ability of a GS method in regard to data simulated with the conventional approach in comparison to our alternative approach, in which the metabolome level is additionally considered. To realize the metabolome level in the alternative approach, a curated metabolic network was used, which allowed the change of kinetic parameters of enzymes of the metabolic network according to the genetic information. This systems biology approach enabled in the alternative approach that genetic effects were implicitly simulated, whereas in the conventional approach genetic effects were explicitly simulated.

- The second task was to investigate the different relationships between milk metabolites and milk traits and to gain a deeper understanding within each system-level. For this purpose, various methods from the field of bioinformatics (e.g., machine learning methods) as well as biostatistical methods (e.g., variance analysis (ANOVA)) were applied. Sets of milk metabolites eligible to predict milk traits were also investigated in order to enable the analysis of milk traits from a metabolic perspective and to shed light on a possible functional background for some of the detected associations.

- The third task was to investigate if an improvement in genetic value prediction can be achieved when the additional intermediate level is used. Hence, an integrative

bioinformatics approach is proposed, which is termed metabolite approach. The metabolite approach consists of three steps:

(a) First, identify the metabolites with the highest impact on an investigated phenotype.

(b) Second, use identified metabolites to select SNPs or to weight SNPs.

(c) Third, use selected or weighted SNPs for the genetic value prediction.

This metabolite approach was used twice, once on simulated data and once on experimental data, to investigate if and how much the prediction power could be improved if only a part of the metabolome is considered. In this context, in the simulation study was also analyzed using the whole simulated metabolome. Studying different relationships between the three system-levels enabled a deeper understanding of the associations of these system-levels. It is also possible to reveal new relevant biological information if the experimental data are investigated.

This thesis is highly interdisciplinary, since different approaches from several research fields were combined; especially bioinformatics, dairy cattle science and an approach of systems biology, working with both simulated and own experimental data, and conducting biostatistical analyses. Hence, in the remainder of this introduction the current state and fundamental knowledge of main research fields are presented, especially regarding their role for this thesis. The introduction is structured as follows: First, basic terms of quantitative genetics and population genetics are defined, which represent the basis for the field of dairy cattle science. Afterwards, the background of dairy cattle breeding will be explained, followed by GS methods. This section is followed by a brief outline of information on the bovine genome, i.e., marker maps, quantitative trait loci (QTL) and the important role of cow's milk, including definitions of some important terms. Finally, the fields of bioinformatics and systems biology, including the used metabolic network, are explained and their roles for this thesis are specified.

1.2 Basics of dairy cattle breeding

In this chapter important definitions and specific terms are introduced that are relevant for the understanding of animal husbandry and which also play an important role for the simulation of data in Chapter 3. Further relevant definitions are explained in the corresponding chapters. All analyses presented in this thesis are based on quantitative genetics as well as population genetics. First, the concept of the population is introduced.

Population: "A population, in genetic sense, is not just a group of individuals, but a breeding group; and the genetics in a population is concerned not only with the genetic constitution of the individuals but also with the transmission of the genes from one

generation to the next generation" (Falconer and Mackay, 1996, p. 2).

Different specific population parameters can be estimated. Hereafter, specific population parameters are presented, which are important for the genetic value prediction as well as for the data simulation in this thesis.

Genotype-phenotype map (GP map) and the genetic variability: In general, "observed phenotypes (P) of a trait of interest can be partitioned, according to biologically plausible nature-nurture models, into a statistical model representing the contribution of the unobserved genotype (G) and unobserved environmental factors (E)" (Visscher et al., 2008). The GP map can symbolically be modeled as:

$$P = G + E. \tag{1.1}$$

The genotype can be differentiated into the following three genetic effect types (Visscher et al., 2008):

- Additive genetic effects (a): each allele (locus) has an impact on the investigated trait;
- Dominance genetic effects (d): intra-locus effects are interactions between alleles at the same locus;
- Epistatic genetic effects (i): inter-locus effects arise from the interaction of several loci, for example, the following inter-locus effects are possible if two loci are considered: aa, dd, ad, da;

and thus the genotype (in Eq. 1.1) can symbolically be modeled as:

$$G = a + d + i. \tag{1.2}$$

In a population the amount of variation of a component (P, G, E) can be measured and is termed variance, whereby the variance is defined "as deviations from the population mean" (Falconer and Mackay, 1996, p. 122). The phenotypic variance σ_p^2 (or total variance) is the variance of the phenotypic values (the observed values) of a trait in a population, and can be expressed as:

$$\begin{aligned} \sigma_p^2 &= \sigma_g^2 + \sigma_e^2 \\ &= \sigma_a^2 + \sigma_d^2 + \sigma_i^2 + \sigma_e^2, \end{aligned} \tag{1.3}$$

where σ_g^2 is the genetic variance and σ_e^2 the residual variance (or non-genetic variance or environmental variance), and it is assumed that both components are independent. Also, σ_g^2 can be partitioned into σ_a^2, the variance of additive genetic effects, σ_d^2, the variance of

dominance genetic effects, and σ_i^2, the variance of epistatic genetic effects (Visscher et al., 2008; Falconer and Mackay, 1996, pp. 122-123). More information about the different values can be found in Falconer and Mackay (1996).
In general, dairy cattle breeding distinguishes between breeding value and genetic value or genotypic value which depends on what genetic effect types are considered. The term "breeding value" is used when only the additive component of the genetic variance is considered. If in addition non-additive genetic effects (dominance and/or epistatic genetic effects) are also considered, then it is called genetic or genotypic value. In this thesis, it will be consistently termed genetic value, except of Section 1.3, whereby the investigated genetic effect types will be specified.

Heritability: The heritability of a trait gives the proportion between the genetic and the phenotypic variation. As an example, if a trait has a $\sigma_g^2 = 0.1$, this would mean that 10% of the variance in the trait phenotype is explained by genetic variation and 90% by environmental influences ($\sigma_e^2 = 0.9$). The heritability is differentiated; if only the additive genetic variation of a population is considered then it is termed narrow-sense heritability (h^2) and can be expressed as (Visscher et al., 2008):

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}. \tag{1.4}$$

If, in addition, non-additive genetic effects are considered, it is termed as broad-sense heritability (H^2) and can be expressed as:

$$H^2 = \frac{\sigma_a^2 + \sigma_d^2 + \sigma_i^2}{\sigma_a^2 + \sigma_d^2 + \sigma_i^2 + \sigma_e^2}. \tag{1.5}$$

The heritability can be divided as follows (according to Weiß et al., 2011, p. 99): values from 0 to 0.2 are considered low heritabilities (e.g., fertility in cattle), values from 0.2 to 0.4 are considered medium (e.g., annual quantity of milk the cow), and values larger than 0.4 are considered high heritabilities (e.g., fat content of milk).

Hereafter, necessary population parameters for the genotype level are introduced which play a role for the pre-processing steps in Chapter 2 as well as for data simulation in Chapter 3. A small example is introduced that is intended to facilitate the understanding of the following definitions. The example population consists of five animals and only one locus A with allele A_1 and the complement allele A_2 is considered. The following genotypes are observed: A_1A_2, A_1A_1, A_2A_2, A_2A_2, A_2A_1.
In this context, the gene frequencies is the "proportions of the different alleles at each locus" (Falconer and Mackay, 1996, p. 2). Based on the presented example, this means that A_1 is an allele at locus A, from this follows that "the gene frequency of A_1, is the proportion [...] of all genes at this locus that are the A_1 allele" (Falconer and Mackay,

1996, p. 2).

Minor allele frequency (MAF): Describes the frequency of the occurrence of the minor allele in a population. In the presented example, we would observe the following allele frequencies: $p = \frac{4}{10}$ is the frequency for allele A_1 and $q = \frac{6}{10}$ is the frequency for allele A_2 on locus A, whereby ten is the sum of all alleles at locus A. In this case, allele A_1 represents the minor allele. In addition, the sum of the frequencies must result in unity ($p + q = 1$) at any locus (Falconer and Mackay, 1996, pp. 1-2).

Polymorphism: A locus is defined as polymorphic if the minor allele occurs with a frequency of more than or equal to 1% in a population. To quantify the amount of the genetic variation at polymorphic loci, the measure of heterozygosity is used. Heterozygosity is the frequency of the heterozygous animals (Falconer and Mackay, 1996, pp. 42-45), in the example population the heterozygosity is $\frac{2}{5}$.

Quantitative trait locus (QTL): A QTL is a segment of a chromosome that has an impact on a quantitative trait such as milk fat content. This DNA segment contains important genes or is linked to genes underlying the investigated quantitative trait (Geldermann, 1975).

Hardy-Weinberg equilibrium (HWE): "A population with constant gene and genotype frequencies is said to be in Hardy-Weinberg equilibrium" (Falconer and Mackay, 1996, p. 5), which based on assumptions of an idealized population (see page 8). This relationship between gene frequencies and genotype frequencies plays an important role in population genetics and quantitative genetics (Falconer and Mackay, 1996). In such a population it is possible to determine the expected frequencies of genotypes based on the allele frequencies as follows:

$$p^2 + 2pq + q^2 = 1, \tag{1.6}$$

with

$$\begin{aligned} p^2 &= E(A_1A_1), \\ 2pq &= E(A_1A_2), \\ q^2 &= E(A_2A_2), \end{aligned}$$

where p is the relative frequency of allele A_1 and q is the relative allele frequency of the complement allele A_2 at locus A. $E(A_1A_1)$ represents the expected frequency of genotype A_1A_1 at locus A, respectively for $E(A_1A_2)$ and $E(A_2A_2)$. It is also possible to determine the allele frequencies based on the gene frequencies. Every deviation means that the locus is not in HWE (Falconer and Mackay, 1996, pp. 5-19).
Implementation using the example, the allele frequencies are the following $p = \frac{4}{10}$ for allele A_1 and $q = \frac{6}{10}$ for allele A_2 and thus $p^2 = 0.16 = E(A_1A_1)$, $2 \cdot pq = 0.48 = E(A_1A_2)$ and

$q^2 = 0.36 = E(A_2A_2)$. The obtained expected genotype frequencies are then multiplied with the number of the observed individuals, in our example five, to calculate the expected number of individuals for each genotype, resulting in $E(A_1A_1) \cdot 5 = 1$, $E(A_1A_2) \cdot 5 = 2$ and $E(A_2A_2) \cdot 5 = 2$. In this example, the observed and expected genotypes agree, which means that the locus is in HWE.
In experimental SNP data, this relationship is used as a quality control measure of the determined SNPs, because deviations from the HWE are an indicator for quality problems in the SNP-genotyping procedure or for other deviations (cf. Ziegler et al., 2008). It is typical to determine the expected number of genotypes and then compare them with the observed genotypes using a statistical test to prove for a deviation from the HWE, whereby the observed P-value must be smaller than the predefined HWE P-value (usually 10^{-4}; e.g., Samani et al., 2007; Ziegler et al., 2008).

Linkage disequilibrium (LD): The term was introduced by Lewontin and Kojima (1960) describing the non-random association of the alleles between two or more loci (Slatkin, 2008). The random association is termed linkage equilibrium (LE). Different definitions exist for the measurement of linkage disequilibrium (LD) (Slatkin, 2008). In this thesis, the definition following Hill and Robertson (1968) for a two-locus model was applied, which can be calculated as follows:

$$r^2 = \frac{(p_{A1B1} \cdot p_{A2B2} - p_{A1B2} \cdot p_{A2B1})^2}{p_{A1} \cdot p_{A2} \cdot p_{B1} \cdot p_{B2}}, \tag{1.7}$$

where p_{A1B1} is the frequency of the haplotype with allele 1 at marker locus A and allele 1 at marker locus B. p_{A1B2} is the frequency of the haplotype with allele 1 at marker locus A and allele 2 at marker locus B, and according definitions for p_{A2B2} and p_{A2B1}. p_{A1} is the frequency of allele 1 at the marker locus A, accordingly for p_{B1}. p_{A2} is the frequency of allele 2 at the marker locus A, accordingly for p_{B2}. The degree of the LD in a population depends on different factors (e.g., Slatkin, 2008) :

- Recombination: is the rearrangement of the genetic material leading to new combinations of alleles and possibly to new characteristics (e.g., Charlesworth, 2009).
- Mutation: is the change of the genetic material. Different kinds of mutations exist, whereby in this thesis only point mutations are considered which are changes of a single alleles or nucleotides. The latter is used in Chapter 3 for simulating populations. In addition, a locus is defined as mutated if the minor allele occurs less than 1% (Falconer and Mackay, 1996, p. 42).
- Genetic drift: is "the process of evolutionary change involving the random sampling of genes from the parental generation to produce the offspring generation, causing the composition of the offspring and parental generations differ", or in short, is the random change of genetic variants in a finite population (Charlesworth, 2009).

- Selection: can be differentiated in natural selection and artificial selection. Natural selection means that animals with a higher fitness or a longer life have a higher chance to reproduce than animals with a low fitness or inadequate survival strategies. Artificial selection means that animals were selected by the breeder according to a desired phenotype (Falconer and Mackay, 1996, pp. 184-185).

The presented factors have an influence on the development or loss of genetic variation within a population over time, i.e., generations, and can be measured as LD. In this context, different population genetic models exist to investigate these factors within a population over time. In the field of GS it is common to use a mutation-drift model (see below; e.g., Meuwissen et al., 2001) to simulate a population, until, for example, an appropriate LD is obtained, as can be found in real cattle populations. In this thesis, two population genetic models (i.e., drift model and mutation-drift model) were used to simulate appropriate populations regarding LD in Chapter 3. Hence, the concept of an idealized population as well as the effective population size is introduced next.

Concepts of idealized population and the effective population size: A population is defined as an idealized population if it complies with the following conditions (according to Falconer and Mackay, 1996, pp. 49-50):

1. A huge population size ($N \to \infty$) is considered.
2. Random mating only within the population (including self-fertilization) is allowed and thus migration is excluded. Random mating describes the mating system within a population; each individual has an equal chance to mate with another individual in the population (Falconer and Mackay, 1996, p. 5).
3. The generations are distinct, meaning new generations do not mate with individuals from a previous generation.
4. The number of breeding individuals are constant over all generations.
5. No selection, and
6. No mutation is allowed.

The idealized population does not reflect the reality of a real population and some of the conditions cannot be met in a real population. If such an idealized population is considered over time, the HWE law comes into play, which states that in this case the gene frequencies and the genotypes are constant from one generation to the next. In addition, the genetic drift has also no impact on the genetic composition.
To describe the effect of genetic drift on the genetic composition within a population, the term effective population size (N_{eff}) was introduced by Wright (1931). This term

represents the rate of change in the genetic composition within a population, i.e., it represents the "[...] random sampling of genetic variants in a finite population" (Charlesworth, 2009) and can be determined as follows:

$$\frac{1}{N_{eff}} \approx \frac{1}{4N_s} + \frac{1}{4N_d}, \tag{1.8}$$

where N_s is the number of sires and N_d the number of dams. If N_s and N_d have equal size, the same properties can be observed as in the Wright-Fisher model. In the Wright-Fisher model, conditions two and three from the definition of an idealized population are assumed (Charlesworth, 2009). In a Wright-Fisher model the genetic composition within a population can change over time. Furthermore, the impact of genetic drift depends on the population size (N). Large populations, for which mutation or selection are not considered, show a behavior similar to that of an idealized population, whereas in small populations the genetic drift has a strong impact (Charlesworth, 2009). In dairy cattle breeding, finite populations typically contain more dams than sires, so N_{eff} is significantly smaller than the population size (cf. Eq. 1.8), because half of the genetic material of the offspring generations comes from few males. In this context, in Holstein populations the N_{eff} is estimated around 100 (e.g., Qanbari et al., 2010). N_{eff} is small due to selection over many years. It is common, as mentioned earlier, to use a mutation-drift model in the field of GS to simulate populations with a finite size over time, for example until an appropriate LD is reached. The mutation-drift model considers mutation and genetic drift using for example an N_{eff} of 100 animals with $N_s = 50$ and $N_d = 50$ and conditions two and three of the idealized population are assumed, excluding self-fertilization. In general, the mutation-drift model is based on a Wright-Fisher model with mutation. Afterwards, some generations are generated to simulate the typical half-sib structure as can be found in real cattle populations. In these generations, mutation is excluded and N_{eff} is set for example $N_{eff} = 1{,}000$ with $N_s = 50$ and $N_d = 950$.

1.3 Background of cattle breeding

This part provides a brief overview of the historical development of dairy cattle breeding, especially the importance of GS.

"Selective breeding has been going on for thousands of years and with increasing intensity during the recent centuries. Such selection, over many generations and in large populations, has driven the accumulation of new mutation with favorable phenotypic effects, as well as the development of alleles and haplotypes that differ by multiple functionally significant substitutions" (Andersson, 2001). In the field of dairy cattle science selection is done to reach specifically defined breeding objectives. The breeding objective is the goal which should be achieved by selection (also termed breeding). In general, breeding is the systematic selection and mating of domestic animals. Among the animals (at the reproductive age) are those animals selected that correspond "best" to the breeding

objective. The breeding objective is set on the one hand by breed associations (union of breeders) and on the other hand by breeding companies. Additionally, each breeder can have further individual breeding objectives (Weiß et al., 2011, p. 102). It is typical that such breeding objectives are mostly designed in respect to several traits (e.g., Dekkers and Gibson, 1998) and also that they depend on several factors; for example, on breed purpose, such as beef or dairy production. The objective of each breeding program is to gain the highest possible genetic progress per unit of time, whereby breeders have three possibilities to influence the genetic gain (Seefried et al., 2010):

1. Accuracy of selection is based on estimated genetic value or breeding value (since only additive genetic effects are considered), which is provided by the artificial insemination. In the 1950's the use of artificial insemination was established (Meyn, 2005). At current state, about 80% of dairy cows and heifers are artificially inseminated in Germany (Weiß et al., 2011, p. 130).

2. Generation interval, which depends on how fast the required information can be collected to estimate the breeding value, e.g., the milk quantity of daughters.

3. Intensity of selection, which depends on the breeding stock and how many breeding animals are necessary to keep the stock. This means that the more offspring of a breeding animal can be expected, the better (Weiß et al., 2011, p. 107). As an example, if we have a breeding stock of 100 animals and only five bulls are needed to keep the stock, that is much better than if 50 bulls are needed. In dairy cattle science, it is common to select strongly for bulls and weakly for cows (Weiß et al., 2011, p. 107), which can also clearly be seen in the conventional progeny testing program, see Figure 1.1.

Selection intensity alone is not decisive for a successful breeding program, accuracy of breeding value estimation also plays a large role (Weiß et al., 2011, p. 108). Furthermore, highly heritable traits are more accurately estimated than those with lower heritability. To estimate the milk performance it is common to fit a linear mixed model, which has the general form (e.g., Kruuk and Hadfield, 2007):

$$y = X\beta + Z_a a + \sum_k Z_k u_k + e, \tag{1.9}$$

where y is a vector which contains the observations of a trait of interest, X is the design matrix and β contains one or more fixed effects; for example, milk test-day (more information for fixed effects is presented in Section 1.5.3). Z_a represents the incidence matrix and a the additive genetic effects. Other random effects (e.g., dominance effects) can be included in u_k with the corresponding incidence matrix Z_k, and e represents the vector of residual effects. The fixed effects include all effects that are not of interest, and correcting for these effects is recommended if they are known. The random effects are

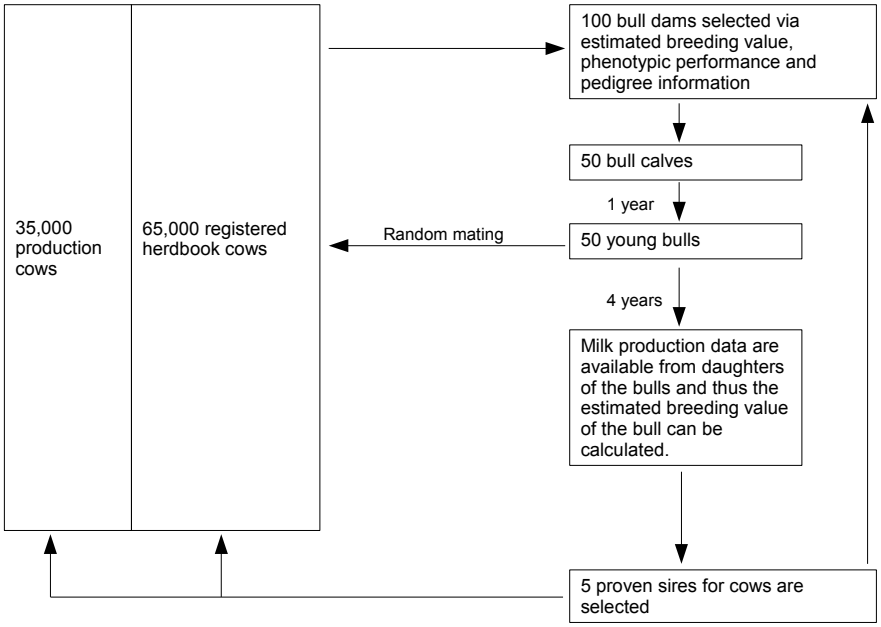


Figure 1.1: Scheme of a conventional progeny testing program for the final selection of 5 cow sires per year. The scheme was created based on König et al. (2009).

those of interest. Different estimation methods are proposed in the literature and a brief historical summary of the development of these estimation methods is presented in the following.

From traditional breeding to genomic selection: Henderson (1949) developed the method which is called best linear unbiased prediction (BLUP). This method allows the simultaneous estimation of fixed effects and breeding values using a mixed model. BLUP is commonly used for domestic animals and it is still used today. Over the years, this method has been further developed. In the first years simple models were used (e.g., sire model), while in recent years more complex models (e.g., animal model considering sires and dams) were applied (Mrode, 2005, p. 39), as the computing power increased over the years. The traditional estimation of breeding values is based on pedigree information and phenotypic data (e.g., Hayes and Goddard, 2010). The accuracy of the traditional estimation of breeding values increases with increasing age of the animals and its relatives as more phenotypic information is acquired.

The development of genetic markers started in the 1980s (e.g., Collard et al., 2005), which provided the basis for using the genome level as third level besides pedigree information and phenotypes for the estimation of breeding values. The genetic markers are used to build linkage maps or genetic maps and physical maps (see Section 1.5.1 for more information). These maps can be used to identify locations on the genome which contain

genes and QTL related to the trait of interest. In general two main strategies exist to detect QTL: association test use candidate genes and genome scans based on linkage mapping with anonymous marker (Andersson, 2001; Mackay, 2001; Ron and Weller, 2007). More information about QTL in Holstein are presented in Section 1.5.2. In general, the 1980's and 1990's can be summarized as the time of linkage maps and QTL hunting (van der Beek, 2007).

Since the early 1990's genetic marker information has been used in dairy cattle breeding schemes (Spelman et al., 2007). Meuwissen and Van Arendonk (1992) presented a framework how the important genome region can be integrated for breeding:

1. Search for the genetic markers within breed and species.
2. Determine the LD between markers.
3. Determine the association between marker and QTL.
4. Embedded the marker information for the breeding program.

The use of informative genetic markers, which are linked or close to known QTL, for breeding programs is termed marker assisted selection (MAS). Three different kinds of genetic marker exist, which are in different relation to QTL and can be used for MAS (Dekkers, 2004). In this thesis only the LD markers, especially SNP-markers, are of interest, whereby LD markers are defined as "[...] loci that are in population-wide linkage disequilibrium with the functional mutation" or "[...] with the QTL" (Dekkers, 2004). The main advantages of MAS in breeding programs were summarized in Mackinnon and Georges (1998). From my point of view the main advantage of MAS is that the genetic information from an animal can be obtained at any time point of its life. This allows breeders to select animals at an earlier state and thus shortened the generation interval. The duration of the generation interval for traditional selection and GS based selection is shown in Figure 1.2 for sire selection. The first estimation method that includes all three kinds of data (phenotype, pedigree and marker information) was proposed by Fernando and Grossman (1989). They proposed an extended version of the traditional BLUP, termed MA-BLUP (MA stands for marker assisted), in which additional genetic marker information is used. In 2003, MAS was introduced in Germany (Bennewitz et al., 2003), and since 2005 MA-BLUP is regularly used for milk performance traits (Reents and Reinhardt, 2007). In this time only few markers were applied for the prediction of breeding values.

Since 2008, SNP panels with around 50,000 SNP markers have been available (e.g., Bovine SNP chip 50K and today SNP panels with 777k are available; Illumina, 2012). Since this time, MAS has been divided into a two-step process (Hayes and Goddard, 2010):

1. The genome-wide SNP markers and trait of interest are analyzed in a genome wide association studies (GWAS).

2. The detected significant SNP markers from the first step are incorporated into the breeding value prediction.

GWAS tests each SNP for a possible association with the trait of interest, which results in a multiple testing problem (Hastie et al., 2009, pp. 683-693). From this follows that SNPs may receive an appropriate error term lower than the predetermined level of significance (also termed false-positives) and thus they are falsely included for the second step, i.e., SNPs are possibly biased for the breeding value prediction (Hayes and Goddard, 2010). GWAS is based on the assumption that a significant association between SNP and trait exists when the SNP is close or linked to a QTL. The dense SNP chips also allow the use of all SNP markers for the breeding value prediction, because it is expected that each QTL is in LD within at least one SNP. This assumption is made in GS. The difference between MAS and GS is: MAS uses only the significant SNPs from GWAS, whereas in GS all available SNPs are used. This implies that in GS there is no need to know QTL positions. GS has also the advantage that the first SNP filtering step is not necessary compared to MAS and thus SNPs are unbiased when they are used for the genetic value prediction (Hayes and Goddard, 2010). The term GS was introduced in 1998 by Visscher and Haley at the 6^{th} World Congress on Genetics Applied to Livestock Production (WCGALP in Armidale; van der Beek, 2007; Meuwissen, 2007). The corresponding analytical framework for GS was presented in 2001 by Meuwissen et al. (2001). In general, the same advantages as in MAS are expected. Since August 2010 the GS breeding values are confirmed and validated by ICAR (International Committee for Animal Recording)/Interbull (sub-committee of ICAR) and determined by the vit Verden in Germany (Reinhardt et al., 2011). In addition, GS also allows the estimation of breeding values without pedigree information, which means phenotypes and marker data are adequate and sufficient for the genetic value prediction (Hayes and Goddard, 2010).

GS can be seen as an independent way to determine breeding values and it complements the traditional breeding. The officially estimated breeding value should include all available information, which includes that the traditional and GS should be done in parallel to reach high accuracy (Seefried et al., 2010). Also a change in breeding programs is expected based on GS. In this context, in near future it will be standard to genotype bull dams and to select them based on the observed genomic breeding value, which can lead to further selection improvements (König et al., 2009).

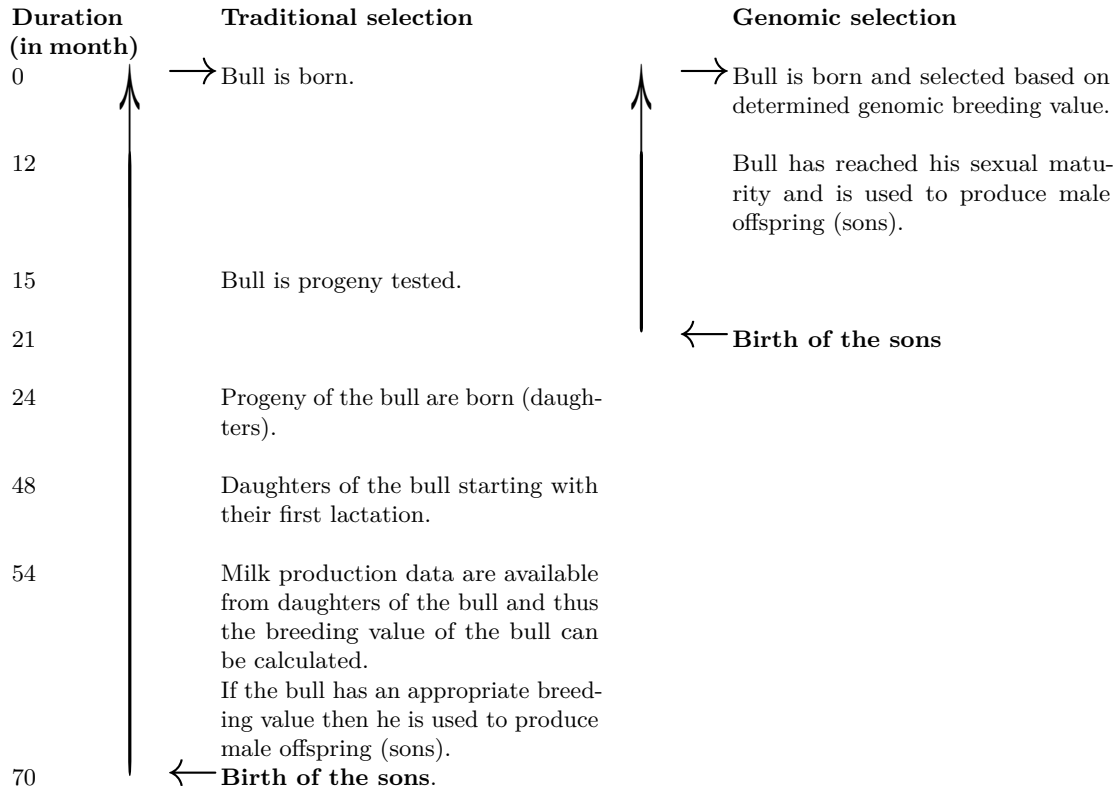


Figure 1.2: Comparison of the duration of artificial insemination breeding programs. On the left hand the traditional breeding program and on the right an aggressive breeding program with the use of genomic bulls. The figure was created following Schefers and Weigl (2012).

1.4 Genomic selection methods

In the field of GS several estimation methods are proposed, for example Bayesian methods, non-parametric and semi-parametric methods (e.g., Dekkers, 2012; Meuwissen et al., 2013). All these methods have the following underlying prediction process in common: Genotypes and phenotypes of the ancestral generation are used to estimate the genetic effects. The estimated genetic effects are used in combination with the genotypes of the offspring generation to predict genetic values. The accuracy of the genetic value prediction can be measured as the correlation between estimated and observed genetic values, which is possible in simulation studies in which the simulated genetic values are known. In experimental data the correlation between estimated genetic values and the observed phenotype values is used.

The main objective of this thesis is to investigate if genetic value prediction can be improved by considering the additional system level of the metabolome; no new estimation methods are proposed, compared or optimized. Two estimation methods were used, both of which in a Bayesian framework. In general, a Bayesian framework allows the inclusion of prior knowledge; prior assumptions mirror the distribution of the genetic effect sizes

of QTL. As the true QTL are typically unknown, priors are defined for SNPs which are in strong LD with the QTL (Hayes and Goddard, 2010). Exemplary settings can be that all SNP effects have an equal impact on the investigated trait, which means all SNPs have small genetic effect sizes (similar to the infinitesimal model of quantitative genetics; Dekkers, 2012; Falconer and Mackay, 1996, p. 438), or that only few SNPs have a moderate genetic impact on the investigated trait and all other SNPs have a genetic effect of zero. To realize the assumption for the underlying genetic architecture, different prior distributions have been proposed in recent literature, e.g., Meuwissen et al. (2001); Verbyla et al. (2010) and for review see Meuwissen et al. (2013). Different approaches for setting the prior are possible (Sorensen and Gianola, 2002); for example, choosing a prior in such a way that the posterior distribution can be solved analytically, or using a simulation to obtain an approximation for the posterior distribution. The latter is mostly done with Markov-Chain Monte-Carlo (MCMC) simulations, for example via a Gibbs sampling algorithm. The Bayesian method BayesB (Meuwissen et al., 2001) can be seen as an appropriate method for our purpose as the literature mostly shows that BayesB delivered high accuracy in genetic value prediction compared to other estimation methods (e.g., Meuwissen et al., 2001; Daetwyler et al., 2010; Habier et al., 2011). BayesB assumes that a large number of SNPs have a genetic effect of zero and a small proportion of SNPs have a moderate genetic effect. Further, the prior distribution of variance, which represents a measure of uncertainty on genetic effects, is a mixture distribution of point mass at zero and an inverse χ^2 distribution. The MCMC technique commonly used is known to be time-consuming. This method estimates all SNP effects simultaneously. Meuwissen et al. (2009) proposed a fast algorithm which makes similar assumptions as BayesB. This algorithm solves the problem analytically and thus avoids the time-consuming MCMC technique. The prior distribution of variances is in this case a mixture distribution with point mass at zero and an exponential distribution. This method is an iterative approach, where the genetic effects are successively estimated for each SNP. Both methods, the fast BayesB as well as BayesB, have been extended to non-additive genetic effects (Wittenburg et al., 2010). The difference in computing time were compared between both methods based on a data set with 2,000 animals in the parent (training) generation as well as in the offspring (test) generation, considering 5,227 SNPs including 23 SNPs that were selected as QTL (Wittenburg et al., 2010). For both estimation methods the used analysis model considers additive as well as dominance genetic effects. The obtained computational time was one second for fast BayesB and four hours using the original BayesB method for one data set using a 2.93 GHz multi-user system (Wittenburg et al., 2011). If pairwise epistatic effects were additionally considered the computational time was seven hours for one data set using fast BayesB. Although fast BayesB requires essentially less computing time than BayesB, it has similar accuracy of genetic value prediction. In this thesis the fast algorithm of the BayesB method was applied as numerous analyses had to be performed using varying numbers of QTL and

SNPs. For analyses a linear model was used, considering the additive and dominance genetic effects (cf. Chapter 3).
In the field of GS it is also of interest to design low-density SNP panels based on the high-density SNP panels, which can be used for a broader screening. Currently, it is common to genotype only elite animals, especially bulls (cf. Section 1.3). To determine such a suitable SNP subset from a high-density SNP panel, different strategies were proposed in recent literature (e.g., Weigel et al., 2009; Verbyla et al., 2009; Moser et al., 2010, more details are presented in Chapter 4). By using such methods, it is possible to select important SNPs that have a genetic impact on the investigated trait of interest. Ishwaran and Rao (2005) proposed a spike and slab variable selection (SVS), which was adopted and validated by Wittenburg and Reinsch (2011) for the genome-wide estimation of SNP effects. SVS is also based on a Bayesian framework. It has similar prior assumptions as BayesB, but it assumes a mixture of two inverse gamma distributions for the variance of uncertainty for each SNP effect, leading to either a very small (but not zero) or reasonably large genetic effect. Further it infers the proportion of non-zero genetic effects which is involved in determining those SNPs with significant genetic effects. SVS was implemented by a Gibbs sampler. This method is used in Chapter 4 once to select SNPs, resulting in an SNP subset, for the classical approach and the metabolite approach. On the other hand, this method is used for the genetic value prediction.

1.5 Basic information on the bovine genome, QTL and cow's milk

1.5.1 Bovine marker maps

The bovine genome has 30 chromosome pairs, consisting of 29 autosomal chromosome pairs and one allosome (sex specific chromosome) pair. The genome consists of around $3 \cdot 10^9$ basepair (bp), which is the physical unit, it has a length of around 30 Morgan (M) or 3,000 centiMorgan (cM), which is the genetic unit (e.g., Kappes et al., 1997).
Two kinds of marker maps exist, which can be distinguished into the genetic marker map and the physical marker map. The genetic marker map uses cM as measure and provides the positions and relative genetic distances between genetic markers along the chromosomes. Distances between two genetic markers are measured in terms of the frequency of recombination (Collard et al., 2005). In this context, if two genetic markers are far away from each other then the probability for a recombination event is high, otherwise the probability is small. Note, the frequency of recombination can not be linearly converted into the frequency of crossing-over events (Collard et al., 2005). To enable conversion between the two kinds of marker maps mapping functions are used. The most common mapping functions are the Haldane mapping function (Haldane, 1919) and the Kosambi mapping function (Kosambi, 1944). The difference between both functions is that the Haldane mapping function assumes no interference between crossing-over events (meaning no crossing-over event is influenced from another), whereas the Kosambi

mapping function assumes interference (Collard et al., 2005). In general, this kind of marker map can be used to determine the possibility of a crossing-over event occurring between two loci. The recombination rate, θ, gives the probability of a crossing-over event in meiosis, where 100 cM corresponds to one crossing-over event (Sturtevant, 1913). The physical map is given in the unit bp, and here the distance between two genetic markers can be exactly determined. Thus, a complete physical map can only be obtained if the whole genome sequence is known (Morton, 2005). In April 2009 the sequence of the bovine genome was completed (The Bovine Genome Sequencing and Analysis Consortium et al., 2009). Both kinds of marker maps cannot be transformed 1:1, because genetic distances and physical distances are not equal and can differ along the chromosomes. The chromosomes contain regions with high recombination frequency, termed "hot spots", and those with low recombination frequency, termed "cold spots" (Collard et al., 2005). Further information about the relationship between the two marker maps can be found in Morton (2005). Both kinds of marker maps are relevant in all chapters of this thesis.

1.5.2 Quantitative trait loci in Holstein dairy cattle

The focus of this section is to roughly explain methods used to detect QTL and quantitative trait nucleotides (QTNs), since this kind of information is used in Chapter 4. The principle of QTL mapping is as follows: "If a QTL is linked to a marker locus, there will be a difference in mean values of the quantitative trait among individuals with different genotypes at the marker locus" (Mackay, 2001). The idea of QTL is not new as it has already been described by Sax (1923). Basically, QTL show different strength of genetic effects or impact on the investigated trait. A QTL with high genetic effects can be detected easier than a QTL with low genetic effects. In general, the smaller the genetic effect size of the QTL the more animals are required to enable its detection. Different designs were proposed in the literature to detect QTL, where the optimal designs regarding maximal statistical power are backcross or F_2 generation of a cross-population between inbreed lines (Ron and Weller, 2007). Since 1995 it is common to use daughter and granddaughter designs to detect QTL in dairy cattle science using linkage mapping. "In these designs, QTL-marker linkage phase varies across families, and all analyses are performed within half-sib families" (Ron and Weller, 2007). With the aid of genetic markers it is possible to determine the most likely location of a QTL by interval mapping and the corresponding confidence interval can be determined by the non-parametric bootstrap (Ron and Weller, 2007). The typical confidence interval for the QTL location spans often tens of genetic map units, resulting in several genes being contained in a QTL (Ron and Weller, 2007). In such studies only few animals are used for genotyping (sires) and only QTL with large genetic effect sizes will be detected (Ron and Weller, 2007). Such a detected QTL cannot be well established in breeding programs, because several markers are contained in a large confidence interval whereby not all of them will be informative for the investigated trait. Hence, it is necessary to identify the specific

polymorphism that is the cause of the observed trait deviations. Such polymorphism is termed QTN (Mackay, 2001; Ron and Weller, 2007). In this context, Ron and Weller (2007) proposed a strategy of how to determine and validate a QTN. As mentioned earlier, high density SNP chips have been available since 2008, allowing new strategies such as GWAS. In this context, Weller and Ron (2011) noted that more significant SNPs with a genetic impact on the investigated trait were detected using SNP chips than using the traditional designs (daughter and granddaughter design). This review further shows that in the study of VanRaden et al. (2009) the observed results from GWAS correspond to results observed from traditional designs. Finally, it was mentioned that to this date only two QTN affecting milk production traits were identified:

1. DGAT1 is known to have an impact on milk fat and milk protein (Grisart et al., 2004; Weller and Ron, 2011), and
2. ABCG2 is known to have an impact on milk protein (Cohen-Zinder et al., 2005).

Detected QTL can be found in databases; for example, in the AnimalGenome database ("http://www.animalgenome.org/cgi-bin/QTLdb/BT/index", Hu et al. (2007)).

1.5.3 Cow's milk properties

In the history of animal husbandry, milk plays an important role. For a long time, desired milk properties (milk production traits), e.g., to yield a high quantity of milk, and high levels of milk fat and milk protein, and morphological traits were the main breeding objective (up to the mid 1990s; Oltenacu and Broom, 2010). The role of cow's milk is manifold today. Besides the importance as nutrition for newborns (immune response), it also plays a central role as staple food for the population, breeding and animal feed in animal production, and it is used as natural resource in the industry (Töpel, 2004, pp. 1-2).
Milk is secreted by the mammary gland of cows and the milk composition depends on several influencing factors. The latter are presented in Figure 1.3.

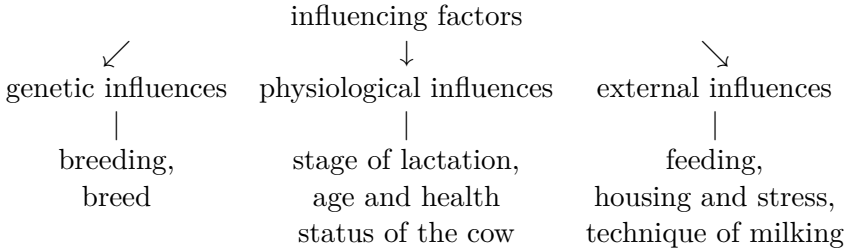


Figure 1.3: Factors influencing milk (adopted from Töpel, 2004, p. 7).

This Figure shows that three main groups of influencing factors exist: genetic, physiological, and external influences. The last two groups are also known as systematic effects.

The various influencing factors imply that it is useful to control the milk quality and quantity in a standard procedure. Quality and quantity of milk is measured once a month in a standardized MPT (Töpel, 2004), which determines milk ingredients quantitatively that are important for the quality of milk and helps monitor animal health. The milk ingredients are usually measured via infrared spectroscopy (ADR, 2008). To enable comparability between cows regarding the obtained milk traits it is necessary to correct for known influencing factors. Such correction is also necessary to allow further statistical tests to obtain, for example, an unbiased estimation of the genetic effects in GS. In the literature different kinds of models (e.g., test-day model, Ptak and Schaeffer, 1993) are proposed to correct for influencing factors (fixed effects). In the field of GS it is common to use mixed models (cf. Eq. 1.9 on page 10) for the estimation of genetic effects within a population. In these models known influencing factors are corrected (e.g., day of lactation and farm). The impact of influencing factors plays a role in Chapter 2, where the specific randomized design is presented, which takes these influencing factors into account for measuring the milk metabolite spectra via gas chromatography-mass spectrometry (GC-MS). The influencing factors also play a role in Chapter 4, where milk metabolites and milk traits are corrected to enable unbiased analyses.
In the following, two terms are introduced which are important in Chapter 4.

Duration of lactation: The duration of lactation is defined as the following time period: one day after calving till cow goes dry. A standardized size in this context is the 305 day milk yield, whereby here the time period is defined between one day after calving until at least day 250th and maximum until day 305th (ADR, 2008). Further, in the first six days the milk is termed colostrum, whereby the milk composition clearly differs from that of the mature milk which follows, and in the last weeks the milk is termed "old milking" milk (Töpel, 2004, p. 8). An idealized general course of the lactation based on the 305 day milk yield is presented in Figure 1.4. In the first period an increase of the milk yield can be observed followed by a slow decrease. Finally a strong drop in milk yield is observed.
Beside milk performance, the energy balance (EB) of the cow is an important term.

Energy balance (EB): The EB can be determined as the difference between consumed energy (feed consumption) and required energy of a cow (Grummer and Rastani, 2003). If a negative EB occurs, the cow utilizes the energy from its fat depot. Otherwise, the cow stores the energy as fat (Grummer and Rastani, 2003). In this context, it is also known that a high producing cow goes through the following phases during lactation (Kirchgessner, 1992, p. 297):

1. In the first third of lactation a negative EB is observed that cannot be completely compensated with concentrated feed.

2. In the second stage of lactation the EB is balanced.

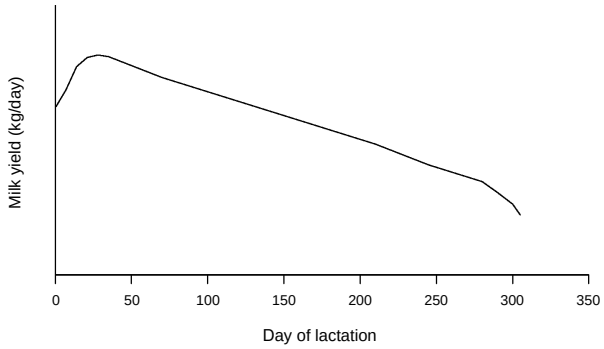


Figure 1.4: Idealized schematic representation of the general course of lactation. The figure was created based on Figure 7.1-7 in Kirchgessner (1992) on page 293.

3. In the last third of lactation a positive EB is observed.

1.6 Bioinformatics and Systems Biology

In this section the relevance of bioinformatics and systems biology to this work is presented. Bioinformatics, as well as systems biology, are interdisciplinary research areas and evolved out of various areas (e.g., Hagen, 2000; Westerhoff and Palsson, 2004). Both have a global view of the system-level information of the GP map. System-level information is provided by different omics data; for example genomics, proteomics, metabolomics. Integration of different omics data allows users to investigate specific associations between different levels as well as to study their interplay to get more insight into the complexity of a biological system (Ge et al., 2003; Westerhoff and Palsson, 2004; Choi and Pavelka, 2012). Both research areas have in common that various definitions exist, whereby some of them are more restrictive than others. Before bioinformatics and systems biology are defined, the term omics is introduced.

1.6.1 Omics

The suffix "omics" is usually used for large-scale data or information in biology. The prefix clarifies which part of biology is considered, such as genomics for the genome, proteomics for the proteome, and metabolomics for the metabolome (Yadav, 2007). In this thesis two different omics levels, i.e., genomics and metabolomics, are considered and will be briefly explained in the following.

Genomics Genomics is termed the research field, where the whole genome, i.e., all genes of an organism are analyzed. The access is gained via DNA sequencing (e.g., Joyce and Palsson, 2006). In Chapter 2 the access of the genetic information of the Holstein Friesian cows is described, which is used in Chapter 4.

Metabolomics Metabolomics is the research field that investigates all metabolites of a biological system. The aim of metabolomics is to identify and quantify the metabolome (e.g., Fiehn, 2002; Krastanov, 2010). Different techniques exist to determine the metabolome and each has advantages and disadvantages (e.g., Roessner and Bowne, 2009; Lei et al., 2011), especially no technique is able to measure all low-molecular substances in a sample (e.g., Roessner and Bowne, 2009; Lei et al., 2011). Mostly a small (biased) fraction of all metabolites is measured experimentally (e.g., Weckwerth, 2003). The most common techniques in metabolomics are nuclear magnetic resonance spectroscopy, liquid chromatography-mass spectroscopy, and gas chromatography-mass spectrometry (e.g., Roessner and Bowne, 2009; Lei et al., 2011). After measuring parts of the metabolome, it is common to apply classical statistical analysis methods, for example ANOVA, and bioinformatics tools, e.g., clustering, for investigating the measured part of the metabolome (Krastanov, 2010). In the field of dairy science, principal component analysis is primarily applied as a first unsupervised analysis method. In other fields, e.g., plant science, different multivariate analysis methods are typically used to explore the data (Sugimoto et al., 2012).
Metabolic data can be used for statistical analyses, for data mining, and for modeling metabolic networks (Fiehn, 2002). In this work, the milk metabolite spectra were measured via GC-MS (cf. Chapter 2). After pre- and post-processing steps, different analysis methods (e.g., multivariate methods) were applied to investigate, for example, the relationship between milk metabolites and milk traits more deeply. All analyses and results for the experimental data are presented in Chapter 4.

1.6.2 Bioinformatics

Numerous definitions for bioinformatics are present in the literature. The most relevant definitions for this thesis is the following. "Bioinformatics [...] is the research field of quantitative analysis of information relating to biological macromolecules with the aid of computers“ (Xiong, 2006, p. 3). More explicitly, "bioinformatics is conceptualizing biology in terms of molecules (in the sense of Physical chemistry) and applying 'informatics techniques' (derived from disciplines such as applied maths, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications“ (Luscombe et al., 2001). Summarizing, "the ultimate goal of bioinformatics is to better understand a living cell and how it functions at the molecular level“ (Xiong, 2006, p. 5).
Bioinformatics can be distinguished into two main fields, the first of which is developing software tools and the construction, maintenance, and curation of databases. The second field involves the application of software tools, where the application can be divided into three application areas: structural (e.g., protein structure prediction), sequence (e.g., genome comparison, sequence database searching), and functional analysis (e.g., metabolic

pathway modeling, protein interaction prediction) to generate biological knowledge. This division does not mean that these kinds of application areas are always considered separately, but rather different areas are integrated to obtain results or find specific connections (Xiong, 2006).

Bioinformatics plays a central role in this thesis and different tools and methods from the field were applied. For example, software tools to create and manage a database were used as well as tools that allow the extraction of desired information from the database (see Chapter 2 for more information) or different bioinformatics techniques, e.g., machine learning approaches and clustering, are applied in Chapter 4 to investigate the milk metabolites and milk traits as well as the relationship between them. In this context, the proposed metabolite approach in this thesis represents an integrative bioinformatics approach, because two different kinds of omics data, i.e., genomics and metabolomics, were integrated. These kinds of data were differently combined in connection with an observed phenotype and analyzed. The following relationships were investigated: metabolome-phenotype, genotype-metabolome, and genotype-phenotype. This integrative bioinformatics approach is used in Chapter 3 and Chapter 4.

1.6.3 Systems Biology

The general focus of systems biology is to understand the underlying components (or structures) and the dynamic behavior of biological systems (Kitano, 2002). The main point of systems biology is to investigate "the behavior and relationships between all of the elements in a particular biological system while it is functioning" (Ideker et al., 2001). In general, large data sets are obtained, whereby a large number of variables (e.g., metabolites) and nonlinear relationships have to be elucidated (e.g., Wolkenhauer et al., 2005). To gain a deeper understanding of the underlying mechanism in the behavior of the complex system it is necessary to make assumptions (abstraction) and use mathematical models from systems theory to describe the structure and the dynamics of a biological system (Wolkenhauer, 2001; Ideker et al., 2001; Wolkenhauer, 2007). In this context, one goal of systems biology is "to turn [...] static maps into dynamic models which can provide insight into the temporal evolution of biochemical reaction networks" (Wolkenhauer et al., 2005). Wolkenhauer et al. (2012) reviewed in this context that till today the main focus in systems biology lies on the reconstruction of biological networks, i.e., gene regulatory networks, signaling networks, and metabolic networks. In general such models are used to get a deeper understanding of the mechanism of the underlying cell function after the models are validated with experimental data (Wolkenhauer et al., 2012). For the design of such biochemical networks (e.g., metabolic network) several types of modeling (or mathematical formalism) are proposed in the literature, for example, qualitative, semi-quantitative, and quantitative models (e.g., De Jong, 2002; Kærn et al., 2003). The latter kind of model is of interest in this thesis, not in respect to design or to optimize an existing model, but rather to use a curated metabolic network (see

below). Inspired by Mendes et al. (2003) the metabolic network is used, where the metabolome level of the classical GP map is additionally considered (Chapter 3), to realize the genotype-metabolome step in our proposed alternative approach to simulate more realistic data. Mendes et al. (2003) simulated different gene expression data sets based on artificial gene regulatory networks. These network models are composed of coupled ordinary differential equations (ODEs), where each equation describes the production and degradation dynamics of a specified gene product. In this context, mostly ODEs are used for quantitative modeling of intracellular dynamic processes to express temporal changes in concentrations (quantities; e.g., Edwards et al., 2002; Wolkenhauer et al., 2005; Polynikis et al., 2009). Further, Mendes et al. (2003) realized biological variation by adding random values to the kinetic parameters. Liu et al. (2008) adopted this approach and followed Mendes et al. (2003) by incorporating QTL variation to influence the kinetic parameters in their gene regulatory network. Based on these two approaches, we make use of an existing curated and parameterized metabolic network model, i.e., SBML model (SBML: Systems Biology Markup Language; Hucka et al., 2003) to obtain different metabolic outcomes in dependency of the genetic information (more information for the used metabolic network is presented in Section 1.6.4.2). Such models can be found in databases for example, in the BioModels Database ("www.biomodels.org", Le Novère et al., 2006).

1.6.4 Metabolic networks

In this section only a brief overview of the wide field of metabolic network modeling is presented. It has been a long tradition to try to understand the cellular regulation of metabolism (Heinrich and Schuster, 1998; Fiehn, 2002). The field of modeling metabolic networks can be distinguished in several sub-fields. An overview of the sub-fields and the different advantages as well as disadvantages can be found in a review from Wiechert (2002). In this thesis only the type of constraint-based modeling of metabolic networks is of interest (Terzer et al., 2009; Ruppin et al., 2010). In particular, the flux-balance analysis (FBA), which is explained in the next section, as the used curated metabolic network is based on FBA.

1.6.4.1 Flux balance analysis of metabolic networks

Usually, not all details of kinetic information are available to reconstruct a cellular metabolism in mathematical detail for a single cell (Bailey, 2001; Edwards et al., 2002; Ruppin et al., 2010). An exception in this context is represented by the human red blood cell where an advanced level of mathematical modeling exists (Schuster and Holzhütter, 1995; Edwards et al., 2002). The core of constraint-based models and thus also for FBA is the stoichiometric matrix (S) of size *metabolites* x *reactions*, which is the mathematical representation of the reactions of a metabolic network; elements of such matrix are termed

stoichiometric coefficients. Rows correspond to the metabolites and columns represent the chemical transformation of each catalyzing enzyme (reactions), which indicates how many molecules of each metabolite are transformed. The stoichiometric matrix represents the metabolic reactions: consumed (negative sign), produced (positive sign), and zero for each metabolite that does not participate in the corresponding metabolic reaction (Orth et al., 2010). Based on this concept the balance equations or mass reactions, v, using ODEs for participating metabolites of the metabolic network can be formulated. "These balances simply state that the concentration change of a metabolite over time is equal to the difference between the rates at which the metabolite is produced and consumed" (Edwards et al., 2002). It is assumed that the metabolic network has reached steady-state ($Sv = 0$; also termed as quasi or pseudo-steady-state, Orth et al., 2010), meaning that the metabolite concentrations do not change (results in metabolite balancing equations). "It requires that each metabolite is consumed in the same quantity as it is produced, and this is the basis for further analysis of metabolic fluxes based on the stoichiometric matrix" (Orth et al., 2010). In realistic large-scale metabolic models, more reactions exist than compounds (metabolites), which results in no unique solution to the ODE system (Orth et al., 2010). To reduce the possibilities of solutions, constraints can be used, which can be formulated mathematically, for example, thermodynamics for the reaction direction and the enzyme kinetic maximum reaction velocity ($Vmax$) for enzyme kinetics (Palsson, 2000; Edwards et al., 2002; Terzer et al., 2009). In general, each constraint has an influence on the solution space and narrows the solution space in which the metabolic network must perform (Palsson, 2000). The core of FBA is to identify such an optimal solution within a constraint space via linear optimization (Palsson, 2000). In the easiest case only a single flux is either maximized or minimized. For FBA different optimization criteria are proposed in the literature; for example, to optimize biomass. It is also possible to integrate experimental measurements as constraints to support the calculation of the entire metabolic flux distribution (Edwards et al., 2002). In general, "FBA can be used to calculate, interpret and predict metabolic flux distributions and to analyze the capabilities of a metabolic network based on the systemic stoichiometric, thermodynamic and reaction capacity constraints" (Edwards et al., 2002).
This overview was presented to give a little insight of FBA, because the metabolic network model used in this thesis is based on FBA.

1.6.4.2 Usage of the curated metabolic network

In Chapter 3 the curated SBML model 70 (Holzhütter, 2004), which was downloaded from "http://biomodels.org/" (Le Novère et al., 2006), was applied to realize the genotype-metabolome step in our alternative simulation approach. This model was curated such that the model simulated the same flux values as given for the "kinetic model" as presented in Table 1 of the original paper from Holzhütter (2004) (adopted from the corresponding SBML model report). In this context, it is mentioned that there is a discrepancy between

the original paper and the SBML model available online. The original model in the published paper is based on 30 reactions and 29 metabolites, whereas the official SBML model is based on 38 reactions and 45 metabolites.
The SBML model belongs to the category of FBA, where minimization of all fluxes is used as optimization criterion (Holzhütter, 2004). Holzhütter (2004) proposed this criterion, because he mentioned that maximization of the biomass is appropriate for primitive cells like bacteria, but a more general criterion for optimization is necessary for cells with "more sophisticated ambitions". The model is based on ODEs, whereby each ODE describes the temporal behavior of metabolite concentrations. The model was designed based on the erythrocyte cycle in human red blood cells. The model comprises the pentose phosphate pathway and glycolysis. In this paper, the author shows that the FBA method with the used optimization criterion provided results which were in a good agreement with those found in the corresponding kinetic model (Schuster and Holzhütter, 1995), even when an enzyme parameter, i.e., $Vmax$, was varied. In this paper, four different fluxes were influenced by varying $Vmax$, in a range between 50% and 500% of the normal values of $Vmax$. The corresponding kinetic model was published earlier by Schuster and Holzhütter (1995), where the metabolic change was investigated for 23 enzymes and the enzyme kinetic parameter $Vmax$ was varied in a range between 0% and 5,000% of the normal values of $Vmax$. In both studies only one enzyme per analysis was investigated. The theoretical predictions were compared with observations from experimental data for several enzyme deficiencies, where good agreements between theoretical and experimental results depicted from the literature were observed.
The SMBL model was selected and considered to be adequate, because among the few existing curated metabolic network models for mammals it belongs to the few larger alternatives, and no curated metabolic SBML model existed for bovids at the time where it was processed (2008/2009).

1.7 Thesis objectives and structure

In the following the objectives and the structure of this thesis is presented. For each chapter the relevant issues of the work are discussed.

In **Chapter 2** the experimental data acquisition and the corresponding pre- and post-processing steps are presented to obtain complete and ready prepared data for the desired three system-levels: genotype, metabolome, and phenotype, for at least 1,300 Holstein Friesian cows. The latter implicates that a huge amount of data had to be obtained during the data collection period, involving several co-operation partners as well as own lab-work. The diverse information had to be obtained at different time points, making a sophisticated experimental design necessary. In this context, the main question was:

2.1 How to organize and ensure data validity and integrity for all the different informa-

tion of the data collection?

It is known that several influencing factors (e.g., milk test-day) have an effect on the investigated milk traits (cf. Section 1.5.3). A similar effect of the influencing factors was expected for the obtained milk metabolites via GC-MS. Hence, milk metabolites should be also corrected for these influencing factors as well as for the expected experimental error, since it is known that over time the column (GC-step) wears out, to enable later unbiased analyses of milk metabolites. This resulted in the following question:

2.2 How to account for known influencing factors of the GC-MS metabolite profiling?

In general Chapter 2 focuses on the realization of storing and digitally archiving the experimental data and the specially created relational database is presented. The developed randomized design is presented, which enables to correct for the influencing factors on milk metabolites after GC-MS measuring and preparation. This part of the work was published in Melzer et al. (2010a).

Chapter 3 contains all information for data simulation and analyses of simulated data. This chapter is consistently divided into three parts.
In general, we wanted to simulate more realistic data using a more complex model of the GP map than it is realized currently in the field of GS (conventional approach). The first part focuses on how to obtain more realistic data with respect to the experimental data. In our approach SNP-genotypes were simulated based on SNP positions from a commercially available SNP chip, which was used for our experimental data. As mentioned earlier an appropriate LD is important for the genetic value prediction to ensure an adequate prediction precision. Therefore a preliminary study was implemented (Melzer et al., 2010b) and in this context the question was:

3.1 Which population genetic parameter settings must be used to achieve a suitable LD in our data sets with respect to experimental data?

In the second part, our alternative simulation approach is presented to simulate more realistic data, using a more complex model of the GP map, i.e., the metabolome level is additionally considered. The simulated data designed to be more realistic data were compared with data simulated using the conventional approach regarding prediction precisions and estimated variance components. Here, the focus was especially on the following questions:

3.2 Is the conventional approach an appropriate basis to simulate data (for methodological development) with similar structure as experimental data?

3.3 If data are simulated using more complex models, is it possible to detect the genetic effects adequately using a linear model?

3.4 How much complexity in the GP map is required for simulated data as a basis to develop methods that are applicable more generally?

Possible deviations from linearity were also characterized in the alternative approach. This part of the chapter was published in Melzer et al. (2013b).
The simulated data of the genotype, metabolome and phenotype level via the alternative approach enabled us to investigate if and to which degree an improvement can be observed if the metabolome is additionally used for the genetic value prediction (Melzer et al., 2011). This is the focus of the third part of this chapter. Our developed integrative bioinformatics approach (metabolite approach) is presented, which is used to analyze the three system-levels. The emphasis of this study was on the following two questions:

3.5 How much gain in prediction precision can be expected if only a part of the metabolome is considered?

3.6 Which measure of weighting is appropriate for important SNPs?

Thus, Chapter 3 focuses on the realization of an alternative simulation approach, including the simulation of data based on the experimental set-up, and the methodological development of an integrative bioinformatics approach to use information of the metabolome for the genetic value prediction.

Chapter 4 contains all analyses involving the experimental data. In this chapter three different investigations are presented and hence it is consistently divided into three parts.
First, a conceptual comparison between simulated data and experimental data was realized to investigate the following question:

4.1 Is the conventional approach an appropriate basis to develop or to optimize estimation methods in the field of GS with respect to the experimental data?

Second, the milk metabolites and milk traits as well as the relationship between both levels were studied in more detail using various analysis methods, e.g., machine learning methods, to enable a deeper insight of these relationships. Furthermore, the effect of the influencing factors on milk metabolites and milk traits were studied. The focus in this part of the work was especially on the following questions:

4.2 Do milk metabolite profiles change during lactation (do they show a lactation curve)?

4.3 How to determine the importance of the measured part of the milk metabolome for the investigated milk traits?

4.4 Do milk metabolites determined to be important play a biological role for an investigated milk trait?

This comprehensive study was published in Melzer et al. (2013a).
Third, the proposed metabolite approach was implemented for the use on the experimental data to investigate the following questions:

4.5 Do SNPs selected for important milk metabolites, which show a high importance for the milk trait prediction, have a relevance for the milk trait?

4.6 How much gain in the genetic value prediction can be observed if the metabolite approach is used compared to the classical approach?

4.7 Does genetic variation on important SNPs (e.g., DGAT1), where it is known that they have a genetic impact on a milk trait, have a similar genetic impact on important metabolites for the investigated milk trait?

Concluding, the third part focuses on the comparison between our metabolite approach and the classical approach based on our experimental data, where especially the important SNPs selected from both approaches were compared with respect to their biological relevance using known QTL as well as QTNs. This study was under review for publication at thesis submission (Melzer et al., 2013c). Summarizing, this chapter contains extensive investigations of the different relationships of the three system-levels of the experimental data.

In **Chapter 5** the conclusions are presented.

Finally, the complete overview of publications resulting from this work is presented and my own contributions are stated.

2 Experimental data acquisition

This chapter deals with the collection of experimental data and the pre- and post-processing steps necessary to obtain the desired system-level data of genotype, metabolome, and phenotype, which are used in Chapter 4. The structure of these data also forms the basis for the data simulation in Chapter 3. The extensive use of the infrastructure of different collaboration partners, different data formats, integrity checks and run-time monitoring of the different levels of data collection was achieved using a relational MySQL database. The GC-MS milk metabolite profiling needs an elaborate experimental design in order to account for influencing factors. The database scheme and GC-MS design presented here may have exemplary character for similar studies. Parts of this chapter were published in Melzer et al. (2010a).

2.1 Introduction

To enable a comparison between the classical approach and the metabolite approach on an experimental level, it is necessary to collect and preprocess the respective data. The design of the experimental data has an effect on the data simulation (in Chapter 3). The data simulation should be conducted in a more realistic way compared to the conventional approach used in the field of GS, where the involved classical GP map for simulating data is based on a simple linear function. In general, we aimed to obtain a complete experimental data set for all three system-levels of genotype, metabolome and phenotype for about 1,300 cows. We collected blood and milk samples. With a sample of blood it is possible to access the genotype after the extraction of the desoxyribonucleic acid (DNA) and determination of the SNP-genotypes. The milk samples were used on the one hand for the MPT, during which different milk traits were measured, e.g., fat content, protein content, and quantity of milk. On the other hand, the milk samples were used to access the metabolome level by using a GC-MS approach (Lisec et al., 2006). The experimental set-up itself generates massive amounts of data. In addition, various information from several co-operating partners had to be stored. As a consequence, different information was obtained at several time points, which makes keeping track of the data during the time of sampling a challenge. Hence, it is necessary to have one central management system and an infrastructure able to handle and organize all this diverse information. For handling and organizing as well as to secure data integrity and validity, a database was built. The database also provides an overview of the collected data at any time point of the collection period.
As post-processing step, it is necessary to correct for influencing factors (e.g., test-day)

for milk traits as mentioned earlier (cf. Section 1.5.3 on page 18) to obtain an unbiased estimation of the genetic effects. To enable the best possible correction for such known influencing factors for the obtained milk metabolites after GC-MS metabolite profiling, a randomized design was created based on the randomized block design (Sachs, 2004, p. 680).

2.2 Data collection

To obtain the information of the genotype, metabolome, and phenotype from the desired 1,300 cows (number of cows based on an advice from the breeding committee at the Fugato status meeting in Potsdam, Germany, in 2008; originally only 500 cows were planned), blood and milk samples from each cow were required. Each cow had to meet the following requirements:

1. The cow had to be in its first lactation period to enable comparability between obtained milk samples, because significant differences exist between different lactation periods (cf. Section 1.5.3 on page 18).

2. The milk sample had to be taken between the 21st and the 120th day of lactation. The time period was chosen for the following reasons: First, in the initial stage of lactation the milk has a significantly different composition (cf. Section 1.5.3 on page 18). To make sure that no cow was still at this initial stage, the earliest time point for taking the milk samples was the 21st day after calving. Second, the time period was limited to ensure a similar status between the cows and avoid significant differences that exist between different stages of lactation. We had to ensure that the interval was not chosen too narrow to collect a sufficient number of samples, but that it was chosen broad enough to keep the desired number of cows.

In total, 1,843 cows from 18 farms within Mecklenburg-Western Pomerania (in the Northeast of Germany) were selected that met the mentioned requirements. Arrangements with farms and the regional institute for standard milk performance testing (Landeskontrollverband für Leistungs- und Qualitätsprüfung, LKV Güstrow, Germany) were made. From each farm we received the ear tag number of the cows that were in their first lactation during the collection year. These ear tag numbers were sent to the Computing Center - IT-solutions for animals (vit Verden, Germany). From the vit Verden we received the corresponding pedigree information (e.g., sires) and the date of calving for each cow. Afterwards, we could start with the sampling of blood (period: January-February and October-November 2009) and milk (March-November 2009). During the period of milk sampling we received weekly an updated list from the vit Verden for all selected cows, e.g, which cows had calved and the day of calving. This list was used to generate a list of cows for every farm for milk sampling. The different lists were provided to the LKV Güstrow via regular updates. The LKV Güstrow forwarded the respective list of the

farm that should be audited within the normal MPT to a milk performance inspector. In addition to normal milk samples, the milk performance inspector also delivered the list of cows and the milk samples specifically taken for our project in a separate box to the LKV Güstrow where the milk samples were analyzed. During the collection period, we received monthly information from the LKV Güstrow regarding the measured values of the MPT for our selected cows as well as from their herd mates. In addition, we had to evaluate the lists of cows returned after being used by the milk performance inspectors. These lists contained information about cows that were sold, medical treatments, or disease status. In this context, I would like to mention that March 2009 constitutes a special case, because for this month we obtained only measured values from the normal MPT of our cows. After March, the milk traits were measured using the extended MPT, which means three additional milk traits were recorded (see Section 2.2.2 on page 34), and we also obtained the measured milk traits information of the herd mates. From this follows that March was more or less a test-month, and we tried to get milk samples from the same cows in the following test-month again whenever possible. Note that MPT compasses the extended MPT in this thesis.

During the period of data collection, we had to cope with processing various types of information, changing data formats, handling missing values, detecting errors data, and react to unforeseeable events. The data collection was conducted during standard business in commercial herds.

The collected samples were further processed to obtain the SNP-genotypes using blood samples and metabolite profiles based on the milk samples. The corresponding processing steps for each system-level are presented in Figure 2.1 and are explained in more detail in the following sections. Note that not all losses of samples are described in detail, as this would go beyond the scope of this work. Nevertheless, some specific numbers are presented to illustrate the dimension.

2.2.1 From blood sample to genotype

The blood samples of all selected cows were collected on a multiple-day tour together with a veterinarian. Afterwards, each sample of blood was divided into four aliquots which were stored at -80°C. To obtain purified DNA-samples from each cow, the blood samples were processed in several steps (cf. Figure 2.1). First, the extraction of the DNA from the blood samples was conducted using the commercially available NucleoSpin BloodL toolkit (Machery-Nagel, Düren, Germany). Before the toolkit was applied to our blood samples, the effectiveness of the toolkit on previously frozen blood samples was verified. For this, the standard protocol of the toolkit was tested on five random blood samples in collaboration with a technician. These were also used for the necessary quality and quantity tests, requested by the Helmholtz Zentrum Munich, which later prepared the SNP-genotypes based on the extracted DNA samples. In this context, the following was requested by the Helmholtz Zentrum Munich: The DNA concentration in a sample should

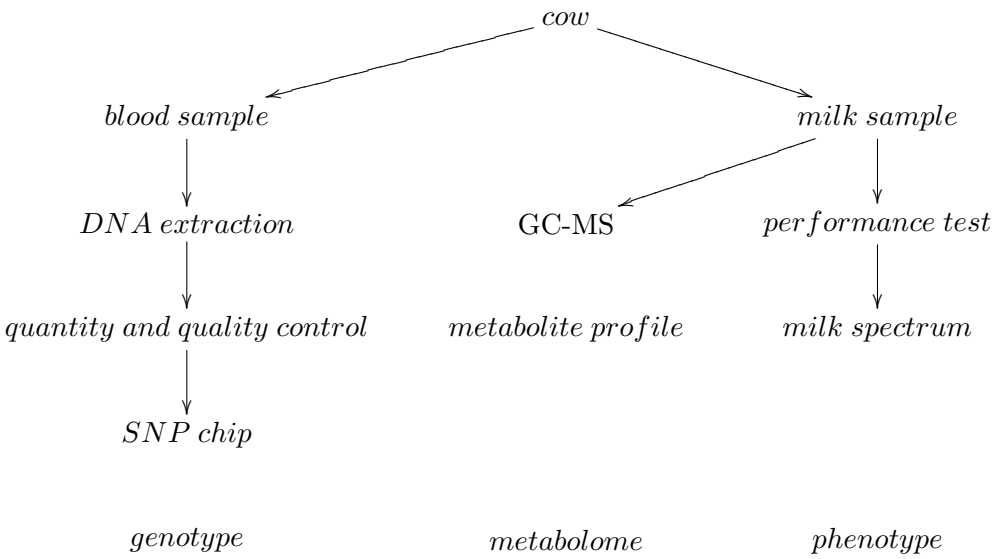


Figure 2.1: Overview of the experimental data preparation to obtain the desired three system-levels.

be in the range of 50-100 $ng/\mu l$ and the quality of a sample should be high enough to enable SNP chip preparation. To ensure these conditions, two standard techniques were applied. On the one hand, the content and the purity of the DNA samples was measured with the NanoDrop 1000 Spectrophotometer (using default setting: DNA-50; Thermo Fisher Scientific, 2008), which measures the concentration of DNA in the unit $ng/\mu l$ and the purity via the ratio of sample absorbance at 260nm and 280nm. The results obtained by the NanoDrop 1000 Spectrophotometer can be saved as csv files; an example graphical output is presented in Figure 2.2 A. The second commonly used technique is the gel electrophoresis, for which the DNA samples, which are charged negatively, and a corresponding DNA marker were loaded on a 1% agarose gel. Then the gel electrophoresis was performed at 100 Volt, whereby the DNA passes from the negatively to the positively charged side. After around 15 minutes, the gel electrophoresis was finished. The gel was photographed under ultra violet light, which makes the DNA visible in the gel, to see how far each DNA sample has run and to compare it to the loaded DNA marker. The gel pictures were used as proof for the Helmholtz Zentrum Munich. An example picture is shown in Figure 2.2 B. After the pre-test was successfully completed, we processed the blood samples of the cows following the standard protocol. First, the quality and quantity check was performed via the NanoDrop 1000 Spectrophotometer. Since the DNA concentration of many samples was higher than the given range, they had to be diluted and then measured again. Afterwards, the gel electrophoresis was conducted. In addition, we knew that 1,344 DNA samples (corresponds to 14 96-well-plates) were possible for the SNP-genotypes preparation. Out of the initially 1,843 selected cows were the DNA extracted from 1,670 cows (without double treatment). For the remaining 173

cows either no blood sample was taken, the sample was incorrectly labeled or no milk sample exist in the desired period of lactation. Further, 326 of 1,670 DNA samples were not included due to either too little extracted DNA or extraction failure, resulting in 1,344 DNA samples. Some of the 326 blood samples were extracted as reserve.

Figure 2.2: An example for each quality and quantity test for DNA. A graphical output from the NanoDrop 1000 Spectrophotometer (A). A picture taken after the gel electrophoresis (B).

After the DNA extraction was completed, the prepared DNA samples (in total 1,344) and the corresponding results from the quantity and quality checks were sent to a laboratory at the Helmholtz Zentrum Munich, where the SNP-genotypes were assessed, using the Illumina® SNP chip 50K (which includes 54,001 SNP positions in total; Illumina, 2008). The obtained SNP-genotypes were further verified using the following steps:

1. The SNPs were identified via BLAST analysis (Altschul et al., 1990) based on the SNP annotation from a physical bovine marker map. The SNP marker map was created by A. Rief. The following steps were realized by Dr. D. Wittenburg.

2. Cows with more than 10% missing SNP-genotypes were excluded.

3. SNP positions which could not be assigned to the corresponding known SNP position of the investigated bovine marker map were skipped.

4. Standard quality checks were applied on the SNP data set (Ziegler et al., 2008): SNPs were excluded if minor allele frequency (MAF) was less than 1%, if HWE was not fulfilled (HWE P-value was set to $< 10^{-4}$; Samani et al., 2007), or if a SNP locus had more than 10% missing values over all cows.

5. The rarely missing SNP-genotypes were imputed using Beagle v3.2 (Browning and Browning, 2007).

During the time of the project the current bovine marker map Btau4.0 has been updated to Btau4.2. To enable analyses based on the latest available bovine marker map, the

presented verification steps of SNP-genotypes were repeated for the new map. In Chapter 3 the simulation studies are presented based on the Btau4.0 map. In Chapter 4, the conceptual comparison between simulated and experimental data is presented based on the Btau4.0 map. For all other experimental analyses in this chapter, the Btau4.2 map was used.
After performing the verification steps, a total of 43,079 SNPs and 1,314 cows were retained using Btau4.0, and 40,317 SNPs and 1,317 cows using Btau4.2.

2.2.2 From milk sample to milk phenotype

The milk samples (one milk sample per cow, the volume per milk sample ranged between 30 and 40 ml) were collected during the standard MPT by a milk performance inspector and delivered to the LKV Güstrow as mentioned earlier. To preserve the milk samples, 150 μl of a 5% sodium azide solution were added per milk sample. Milk samples were analyzed via infrared spectroscopy (Kombi-FOSS, FT6000-FC, FOSS, Hillerød, Denmark). The following traditional milk traits were recorded, whereby the additionally measured milk traits of the extented MPT are presented in bold:

- **Acetone** (%)
- Casein (%)
- Fat (%)
- Lactose (%)
- pH value
- Protein (%)
- Quantity of milk (kg)
- Somatic cell count (SCC) (1000/ml)
- **Saturated fatty acids (SFA)**
- **Unsaturated fatty acids (UFA)**
- Urea (%)

After a milk sample was measured, it was split into four 2 ml tubes (Eppendorf, Germany) and one collection tube with 15 ml (neoCulture-Zentrifugenröhrchen PP, konisch (C8017), neoLab, Rischerstr. 7-9, 69123 Heidelberg) and stored in liquid nitrogen.
After finishing the DNA extraction (without verification steps) and milk traits assessment, we had data regarding the genotype and corresponding milk traits from 1,344 cows in total. In addition, a test run via GC-MS with some of our milk samples had been conducted, from which two milk samples were excluded because they were acidic. On this basis the randomized block design was created for the GC-MS metabolite profiling of 1,342 milk samples. The randomized block design is described in Section 2.3. In the following section, the GC-MS metabolite profiling of milk samples is explained, which was mainly adopted from Melzer et al. (2013a).

2.2.3 From milk sample to milk metabolite profile

After the collection of the milk samples was finished, the samples (one tube per cow) were sent to the Max Planck Institute of Molecular Plant Physiology (Postsdam-Golm, Germany) where the milk metabolite spectra were obtained using GC-MS of the water-soluble phase of each milk sample. For this 100 μl of each milk sample were used (Lisec et al., 2006). The milk samples of 1,342 cows were tested as far as possible according to the predetermined schedule (see Section 2.3 for more information). During the preparation of milk metabolite profiles the predetermined schedule was slightly changed due to laboratory restrictions, resulting in 47 GC-MS batches, not 34 as planned. Most of the predetermined schedule was maintained. For each milk sample the following terms were recorded using the GC-MS: the molecule retention time (GC step) and the mass to charge ratios and the corresponding intensities of molecule fragments (MS step). The obtained molecule spectra were further processed using the R package TargetSearch (Cuadros-Inostroza et al., 2009; R Development Core Team, 2011). The retention time of each type of molecule in a milk sample was converted into a retention index based on the retention time standards of fatty acid methyl esters (FAMEs), which were added to each sample during the GC step. In Figure 2.3 the mass to charge ratios of molecule fragments for the FAMEs are presented over all measured milk samples. In addition it can be seen that the used GC-MS column wears out over the GC-MS batches. Groups of correlating

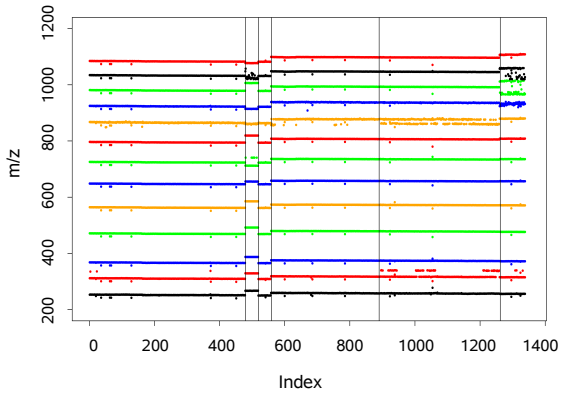


Figure 2.3: The mass to charge ratios (m/z) of molecule fragments for the FAMEs presented over all measured milk samples.

molecule peaks were determined in order to build a metabolite spectrum. For each metabolite spectrum, the median over all samples was determined at every peak. The spectrum of median peaks was compared to reference spectra from the Golm Metabolome Database (GMD; Kopka et al., 2005, http://gmd.mpimp-golm.mpg.de/search.aspx). A metabolite spectrum was accepted or labeled if the similarity index between median metabolite spectrum and reference spectrum was larger than 500 (possible range 0 to 1,000; Cuadros-Inostroza et al., 2009), an example for a metabolite spectrum and the

corresponding reference spectrum is presented in Figure 2.4. If that was not the case, the metabolite was labeled as "unknown" with a specific number. A metabolite was considered for further analyses, if it occurred in more than 80% of the samples, and the log_2 intensity at the largest peak was recorded as an individual observation. This resulted in 187 metabolites for which a reference was found in the GMD, and three unknown metabolites, for which no reference could be assigned. Finally, the nearest neighbor imputation as implemented in the R package pcaMethods was applied to impute missing observations, 7% in total (Stacklies et al., 2007). The obtained data matrix was used for a variety of statistical analyses (in Chapter 4) as recommended by Schwender (2009). The processing of the molecule spectra resulted in 190 milk metabolites for 1,338 cows. All obtained milk metabolites can be found in the Appendix B.1 on page 143. The milk metabolites were further classified into specific active ingredients, for example alcohol and sugar, using the GMD. In Table 2.1 the different classified groups and the corresponding number of metabolites are listed.

2.2.4 Results of experimental data preparation

After all preparation and verification steps were successfully completed for the three system-levels, the different kinds of data were merged together with regard to the cows, whereby a final check was conducted to obtain a complete data set. This was done twice, once for each of the two different bovine marker maps used. As a result, we obtained the complete information for genotype, metabolome, and phenotype for 1,307 Holstein Frisian cows using the Btau4.0 map and for 1,305 Holstein Frisian cows using the Btau4.2 map. Both data sets have in common: 11 milk traits and 190 milk metabolites. More than 2,000 metabolites are expected to exist in cow's milk (Töpel, 2004, p. 3), whereby the 190 milk metabolites represent around 10% of the expected metabolites in milk. The

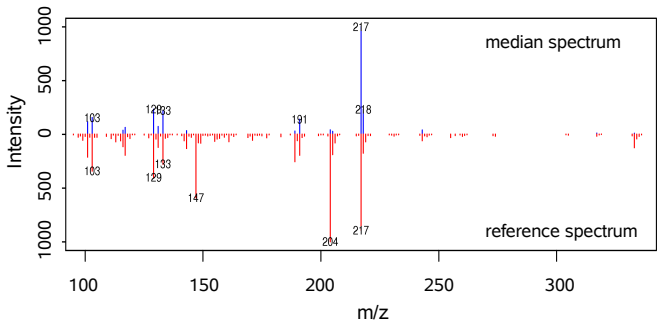


Figure 2.4: An example metabolite spectrum (1,6-anhydro-beta-Glucose); the metabolite spectrum above the null line represents the median of peak intensities from all cows along the mass to charge ratio (m/z) of molecule fragments, and the corresponding GMD reference spectrum below the null line was obtained from a reference substance. The similarity score between metabolite and reference spectrum is in this case 639 from 1,000 possible.

Table 2.1: Overview of the substance classification of the obtained metabolites.

Substance classification	Number of metabolites
Alcohol	5
Aldehyde	1
Amine	3
Amino acid	18
Carboxylic acid	19
Conjugate	2
Indole	1
Lactam	3
Nucleoside	3
Nucleotide	2
Other acid	16
Polyol	5
Purine	2
Pyrimidine	4
Sugar	18
Terpenoid	1
Unspecified	87

genome is covered by 43,079 SNPs using Btau4.0 map and 40,317 SNPs using Btau4.2 map.

2.3 Randomization design for milk metabolite profiling

The randomization design for GC-MS metabolite profiling was created to allow later the correction of influencing factors. Before the realization of our randomized design is presented, the general meaning of a randomized block design will be briefly described. A randomized block design enables the following (based on Sachs, 2004, p. 680):

1. The unbiased estimations of the influencing factor of interest.
2. The unbiased estimations of the experimental errors.
3. An improved normality of the data.

Undesired and unknown correlations are destroyed and thus independent experimental errors are obtained, enabling the application of standard significance tests (Sachs, 2004, p. 680).

At the time point of data collection and preparation, we had 1,342 labeled milk samples as well as corresponding DNA samples. For each milk sample the following influencing factors were known: sire, test-day (milk sampling day), corresponding test-month (in total seven months), and farm. It is possible to measure 40 milk samples per day via GC-MS, thus we had to plan for milk metabolite profiling on 34 days, i.e., 34 GC-MS batches. The GC column, which is used in the GC step of the GC-MS to separate

the substances of a milk sample, wears out over time, and represents an example of an experimental error. This fact implies that influencing factors should be measured on consecutive days, because this allows to distinguish between influencing factors and experimental errors. Based on this knowledge, it was possible to decide which conditions should be fulfilled in the randomization design. The following conditions were used:

1. Each farm occurs at each GC-MS batch.
2. Two test-months at each GC-MS batch.
3. Half-sibs (sires) should be measured during consecutive GC-MS batches.

The order of conditions also represent their importance for the randomization design. This was necessary, because the experimental data are strongly unbalanced as shown in the following. In total, we have seven test-months (May to November 2009) of milk sampling. The quantity of milk samples is imbalanced over the months, on average 192 milk samples per month (range between 9 and 505), and varies strongly as shown in Figure 2.5 A. The frequency of sires also varies strongly, in total 215 different sires were observed, whereby on average each sire had six daughters (range between 1 and 116). The number of cows per farm is unbalanced as well, whereby on average each farm had 75 cows (range between 38 and 138).
To obtain a better result for condition two, the seven milk sampling test-months were grouped into four time points as follows:

- time-point 1: May and June,
- time-point 2: July and August,
- time-point 3: September and October and
- time-point 4: November.

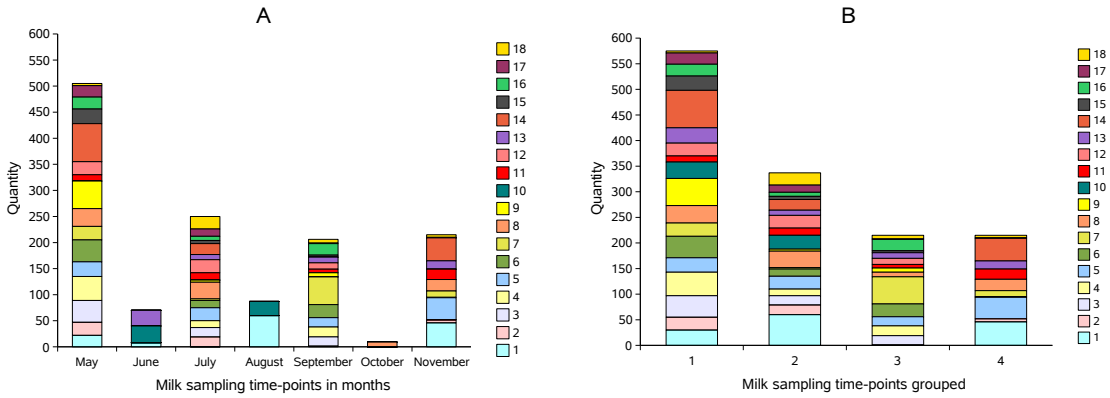


Figure 2.5: Overview of the number of collected milk samples per test-month (A) and number of milk samples after grouping the test-months (B).

The result of grouping is presented in Figure 2.5 B, which shows that an average of 335.50 milk samples per time point were obtained (in a range between 215 and 575 samples). To obtain a design in which all conditions are considered as best as possible we tried to create a randomized GC-MS design which is as balanced as possible regarding the mentioned conditions. In Figure 2.6 the schematic realization of the randomized GC-MS design is presented. Finally, a part of the GC-MS design is presented in Table 2.2. For the realization of the randomized GC-MS design an R-script was written.

Table 2.2: Part of the milk metabolite measuring design. Sires are colored to show that half-sibs were measured on consecutive GC-MS batches.

	GC-MS batch 1			**GC-MS batch 2**			**GC-MS batch 3**		
Farm	Farm	Time-point	Sire	Farm	Time-point	Sire	Farm	Time-point	Sire
7	7	1	34	7	3	34	7	1	81
7	7	2	3	7	4	3	7	3	94
8	8	1	4	8	3	21	8	1	161
8	8	2	107	8	4	188	8	3	3
9	9	1	5	9	3	5	9	1	78
9	1	2	109	7	1	3	7	3	3
10	10	1	12	10	2	3	10	1	25
10	10	2	160	1	2	109	7	3	116

2.4 BovIBI database

During data collection a MySQL database (open source relational database management system), named after the title of the project (BovIBI - Bovine Integrative BioInformatics for genomic selection) was built and expanded according to the additional incoming information. Furthermore, phpMyAdmin (a free software tool; The phpMyAdmin Project, 2006) was used to handle the administration of MySQL in a graphical user interface, which allows users an easy creation of tables, the import and export of tables in various formats (e.g., csv format), and queries to the database. The database is an instrument to monitor, allow common access, i.e., to share all information within the working group, and facilitate handling and checking of the available information. Likewise, it simplifies the processing of data, and it is possible to connect it to databases of other programming and analysis softwares such as R (R Development Core Team, 2008).
As mentioned earlier, we obtained various information at different time points during the data collection. Before new information was uploaded to the database, all information was preprocessed and checked. For example, we received a zip-file with all milk measurements for all cows (selected and herd mates) of the corresponding farms from the LKV Güstrow every test-month. In total four different lists were included for each farm. These lists were merged together in a special order to obtain the desired milk measurements and to extract additional included information. In addition, all measurements with the NanoDrop 1000 Spectrophotometer were stored in the database (cf. Section 2.2.1 on

page 31). To validate or to check all the available information in the database, we used the R package RMySQL (James and DebRoy, 2006), which allows us to extract the desired information from the database via structured query language (SQL, database language) statements. The obtained database results were converted into R variables, which enabled further analyses, e.g., the determination of the dilution of DNA samples. The lists of cows for milk sampling were also prepared in this way and then uploaded, using the R package R2HTML (Lecoutre, 2003) and made available online for the LKV Güstrow and its milk performance inspectors, so that they knew which animals had to be sampled additionally.

In Figure 2.7 the schematic representation of the final BovIBI database is shown. In general the database is based on 19 tables and mainly structured into four parts. The first part contains all information about farms (green), the second part contains all information about the cows (e.g., sold, calving) obtained by the vit Verden (yellow), the third part covers the DNA extraction and SNP chip preparation (red), and the last part contains all information about milk samples and milk metabolite profiles (purple).

Throughout the whole data collection it was necessary to know which data in each part of the database were complete or erroneous and for how many cows all desired information had been obtained. This was necessary to reach the desired number of 1,300 cows, because during data collection we realized that further cows should be selected to obtain the desired target. This resulted in a further collection period of blood samples (October and November 2009) as well as milk samples. Without a database it would be a difficult task to keep the overview over the different kinds of data, to find errors, or make plausibility checks.

2.5 Summary

In this chapter our data sampling procedure and the designed database in which we store the collected data were described, as well as the sophisticated milk randomization design was presented.

Our experiences and the experimental data obtained after the collecting phase show that it is necessary to always have an overview of the data during collection. A database is a helpful instrument to handle and organize data in various ways. Also, the database simplifies access to the data and makes it easy to obtain various information during ongoing data collection. In this context a further aspect which can be summarized regarding the design for an experiment involving large data collection procedures: in our case we could obtain all the desired information for around 71% of the originally 1,834 selected cows. In general, the number of individuals that has to be selected in order to obtain the desired complete data set of an experiment depends on the duration and the type of experiment conducted.

Figure 2.6: Schematic representation of the milk randomization design.
Example assumptions: seven farms, 12 GC-MS batches and 190 milk samples for which the corresponding influencing factors are strongly unbalanced. 'L' are rows which are filled at the end with the leftovers and only used in Step 6.
Step 1: First column and corresponding odd rows are filled depending on farm and test-month, whereby each sire occurs once in this column.
Step 2: The following columns and respective odd rows are filled as follows: For the current column a) fill with samples for which all three conditions are fulfilled depending on the sires from the previous column, whereby each sire is observed once. b) in the other case, samples are used which fulfill conditions one and two.
Step 3: Columns for uneven rows are filled similarly to the second step, whereby now the previous and the following column are used to prove condition three.
Step 4: If a farm has no entry on a GC-MS batch: a) check if samples are available which fulfill condition one and two, and b) check if samples are available which fulfill condition one, whereby only one entry is added to the respective GC-MS batch.
Step 5: For still empty entries prove: a) is a sample available which fulfills only condition 2 and b) random fill.
Step 6: For the last two rows: prove if samples are available which fulfill a) condition two and three, b) condition two, c) condition three and d) fill randomly.

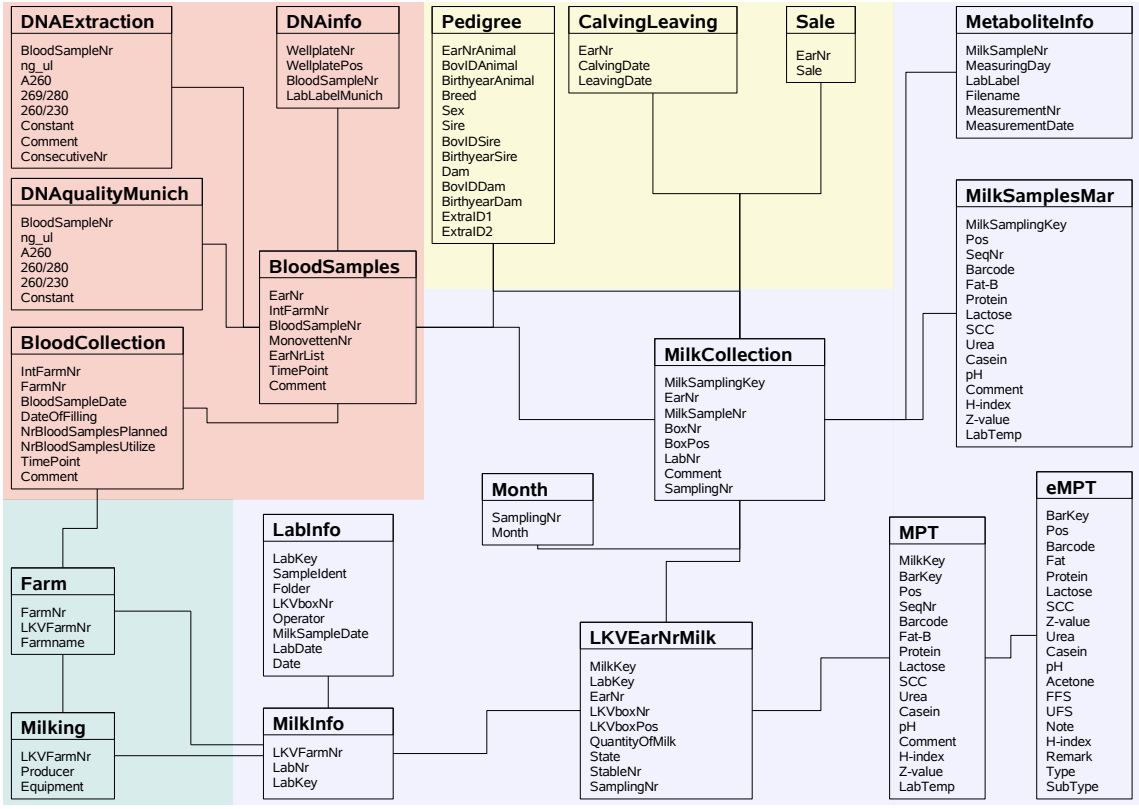


Figure 2.7: Schematic representation of the BovIBI database.

An approach of how to design a GC-MS metabolite profiling taking into account a number of influencing factors for our complex study was presented. The resulting advantage of using the randomized design regarding the statistical modeling to correct known influencing factors after GC-MS metabolite profiling is presented in Chapter 4.

To summarize, we obtained experimental data for genotype, metabolome, and phenotype for 1,305 and 1,307 cows. In the following I do not distinguish between the two bovine marker maps used. Each genome is covered by more than 40,000 SNPs. In total, 190 milk metabolites are available, three of which were unknown. These 190 milk metabolites represent around 10% of the expected metabolites in milk. For the phenotype level, 11 milk traits were measured and the following influencing factors are known: sire, test-day, lactation day, farm, and GC-MS batch.

3 Analyses of simulated SNP-, metabolome-, and phenotype data

The chapter is consistently structured in three parts, whereby different aspects of data simulation are in focus and three different kinds of analyses are presented.
First, in contrast to earlier studies, we wanted to simulate more realistic data sets based on the used experimental SNP chip annotation (cf. Chapter 2). Hence, we conducted a preliminary study to find appropriate population genetic parameters to obtain a more or less realistic LD between SNPs within our simulated populations. On this basis, populations were generated for our alternative approach and the conventional approach.
Second, our alternative approach is presented, striving to simulate more realistic data based on a GP map which includes a simulated metabolome level. The metabolome level was used to simulate genetic values, implicitly including additive and non-additive genetic effects, whereas in a conventional approach additive and dominance effects were explicitly simulated and assembled to genetic values. For both simulation approaches, different scenarios regarding numbers of QTL and SNPs were analyzed using the fast algorithm of BayesB as prediction method. Our analyses revealed that our alternative map showed a smaller prediction precisions (at least 3.75%) compared to the conventional approach in all investigated scenarios. We also observed that the degree of linearity in data simulated with the alternative approach was less (at least 5.88%) compared to the conventional approach. Parts of this chapter have been published in Melzer et al. (2013b).
Third, the alternative approach offers the opportunity to study the influence on the genetic value prediction if the metabolome level is additionally considered. To enable the analysis of the three system-levels an integrative bioinformatics approach, i.e., metabolite approach, was developed with respect to the experimental data (Chapter 4). Different scenarios were investigated regarding using the whole simulated metabolome or just a part of it for the genetic value prediction. Our results revealed, that it is possible to improve the genetic value prediction when the metabolome level is additionally considered.

3.1 Introduction

The introduction is structured in two parts. The first part contains further relevant background information, especially on the state of the art of data simulation in the field of GS. Furthermore the difference of simulating statistical and biological epistasis is illuminated. The second part focuses on our realization of data simulation.

3.1.1 State of the art of data simulation

In the field of GS, data are frequently simulated to compare different methods of genetic evaluation and to optimize methods. These studies have in common that the involved GP map is based on a simple linear function. Generally, it is not known to which degree simulated data following the conventional approach realistically mirror the biology of real traits and if they are sufficient for the development of methods for genetic value prediction. In GS, it is common to simulate SNP-genotypes involving several hundreds or thousands of generations using a mutation-drift model, which leads to a more or less realistic LD between the simulated marker loci. This is usually applied to equally sized chromosomes (e.g., Meuwissen et al., 2001; Calus and Veerkamp, 2007; Habier et al., 2007). Calus et al. (2008) have shown that different spacing of markers has an influence on LD, which in turn has an impact on the precision of genetic value prediction. From the literature it can be concluded that an LD of at least 0.2 is needed to obtain an adequate accuracy in single marker analyses in dairy cattle (Calus et al., 2008). After the initial generations, some populations are generated to obtain the common half-sib structure on which GS methods (cf. Section 1.4 on page 14) then are applied. In this context, various types of genetic effects are discussed to simulate a genetic value considering an additive and/or non-additive (dominance and/or epistasis) mode of gene action (Long et al., 2010; Ober et al., 2011). Note, epistasis is considered in a statistical sense in these contributions, based on the definition of Fisher (1918). In this context, Hill et al. (2008) reviewed that findings based on experimental data seem to point to prevailing importance of additive genetic variance, explaining more than 50% and in most cases close to 100% of the genetic variance. Molecular biology, however, proved that gene action is organized in interactive pathways, regulatory networks, which imply non-additive gene interactions, and, probably, non-additive GP mapping (Moore, 2005). Here, epistasis is considered in the biological sense (Cordell, 2002). In general, the importance of epistasis for mechanisms that underlie the GP map is not yet known (Moore, 2005; Carlborg et al., 2006). It is suspected, however, that epistatic mechanisms may account for much of the causal genetic determination currently unexplained (e.g., Zuk et al., 2012).
In the research field of systems biology, in particularly in its subareas, e.g., for research on gene regulatory networks or metabolic networks (cf. Section 1.6.3 on page 22), epistasis is modeled in a biological manner. As an example, Mendes et al. (2003) simulated different gene expression data sets based on artificial gene regulatory networks. These network models are composed of coupled ODEs, where each equation describes the production and degradation dynamics of a specified gene product. Biological variation is realized by adding random values to the kinetic parameters.

Regarding the choice of model for the genetic value prediction, methods known from the field of GS include genetic effects modeled with a purely additive model, e.g., Meuwissen et al. (2001); Daetwyler et al. (2010); Zhang et al. (2010). However, Lee et al. (2008)

as well as Toro and Varona (2010) have shown that the prediction precision of genetic values increased if an additive-dominance model is used compared to a purely additive model. It has become more and more common to extend existing GS methods to include non-additive genetic effects or to use non-parametric methods (e.g., Long et al., 2010; Ober et al., 2011).

3.1.2 Implementation of data simulation

For our studies, we wanted to obtain a more realistic data set in simulations, which also holds to a conceptual comparison to the experimental data. Thus, we implemented actual lengths of chromosomes and used SNP marker positions from the experimental SNP chip annotation of the bovine genome (cf. Section 2.2.1 on page 31). To obtain an appropriate LD for the initial generation, different simulation scenarios were evaluated in a preliminary study. Subsequently, the populations were generated with the typical half-sib structure, which represent the basis for data simulation.

We drafted an alternative simulation approach designed to be more realistic with respect to the complexity of the GP map: a simulated metabolome level is integrated on top of the classical GP map, whereby the simulation of the genetic effects should also be more realistic, i.e., to model epistasis in a biological manner. Towards this objective, we adopt an approach from the field of systems biology. Mendes et al. (2003) inspired us to model a metabolite level, determining enzyme parameters by marker status at specified marker positions. Liu et al. (2008) followed Mendes et al. (2003) by incorporating QTL variation to influence kinetic parameters in their gene regulatory network (cf. Section 1.6.3 on page 22). Based on these two approaches, we make use of a curated and already parameterized SBML model of the central carbohydrate metabolism (cf. Section 1.6.4.2 on page 24, Holzhütter, 2004), which contains enzymes also found in *Bos taurus*, to realize our simulated metabolome level (download from "http://biomodels.org/", Le Novère et al., 2006), in the following termed SBML approach. Our SBML approach allows us to investigate a more complex GP map, considering an additional level of gene expression in a broader sense. Additive and non-additive genetic effects were implicitly simulated. That means that varying one parameter of an enzyme had an effect on the interactions within the simulated system, which in turn affected diverse metabolite concentrations and not only those catalyzed by the respective enzyme. This offers the opportunity to investigate to which extent a change on the genotypic level leads to a different outcome in the metabolic level. We compare our SBML approach with the conventional approach, where the additive and dominance genetic effects were explicitly simulated. Analyses were realized using the extended fast algorithm of BayesB (fastBayesB; Wittenburg et al., 2011, cf. Section 1.4 on page 14), which models additive and dominance genetic effects. The simulated data using the SBML approach enabled us further to study if an improvement in the genetic value prediction can be achieved when the metabolome level is additionally considered. Hence, we propose an integrative bioinformatics approach, i.e.,

metabolite approach, which allows an analysis of all three systems-levels. The metabolite approach is divided in three steps. First, the metabolite profiles are used to predict a phenotype by applying regression methods from the field of machine learning. Several methods (e.g., random forest, Breiman, 2001) exist which can be used in combination with OMICs profiles (Zhang et al., 2010). From the trained model of the machine learning step it is possible to extract the importance of each variable (metabolite) on the phenotype. Second, the obtained importance values for variables (metabolites) can be used as weights for the simulated QTL, as it is possible to assign the catalyzed metabolites to corresponding enzymes and in turn to assign enzymes to the corresponding QTL. Third, weights and SNP data are jointly used to predict the genetic value using fastBayesB. This resulted in a further study, in which we tested if and to what degree the metabolite approach led to an improvement of the genetic value prediction. Furthermore the influence on the genetic value prediction was investigated when only a part of the simulated metabolome is considered, especially in view of our experimental data set (10% of the expected milk metabolome measured; cf. Section 2.2.4 on page 36).

3.2 Material and Methods

3.2.1 Simulation approaches to obtain a suitable LD

In this section, a strategy for testing the influence of known factors (e.g., mutation or recombination) on the extent of LD in a finite population is presented. In this set-up, the genome is constructed based on available SNP chip annotations.

3.2.1.1 Construction of the simulated genome

A bovine genome-wide SNP data set was modeled in the style of Illumina® Bovine 50K SNP chip (as it was used for the experimental data, cf. Section 2.2.1 on page 31), which was also realized to keep the framework conditions as similar as possible to enable later a purely conceptual comparison between simulated and experimental data (cf. Chapter 4 for more information). From the SNP chip, we used all SNPs with annotated position (bp) according to Btau4.0 resulting in 52,276 SNPs. Chromosome lengths were retrieved from the database Ensembl cow (Ensembl, 2008) to check the plausibility of SNP positions. Three SNPs were omitted because they were outside the corresponding chromosome. SNP positions were linearly converted from the physical map (bp) to the genetic map (cM) using a chromosome-wise scaling factor based on chromosome lengths in cM from the database “marc-USDA cattle” (United States Department of Agriculture, 2008). In Table 3.1 the used lengths of the chromosomes are listed. The conversion from the physical into the genetic map is necessary, because on the basis of genetic distances it is possible to determine the recombination rate (θ) between two adjacent SNPs. The recombination rate gives the probability of a crossover in meiosis (100 cM corresponds to one crossover; Sturtevant, 1913). The recombination rate between two SNPs was

determined using the Haldane mapping function (Haldane, 1919), which is often used in the field of GS (e.g., Meuwissen et al., 2001; Habier et al., 2009):

$$\theta = \frac{1}{2}(1 - e^{-2 \cdot dist}) \qquad \theta \in [0{,}0.5], \tag{3.1}$$

where *dist* represents the distance between two adjacent SNPs in cM.

3.2.1.2 The investigated population genetic models

In this thesis, two population genetic models were applied using different settings to study the influence of known factors (e.g., mutation, genetic drift) on the development of

Table 3.1: Used chromosome lengths for the simulation study for all 30 chromosomes of *Bos taurus*.

Chromosome	Length in bp	Length in cM
1	161,106,243	154
2	140,800,416	128
3	127,923,604	128
4	124,454,208	119
5	125,847,759	135
6	122,561,022	134
7	112,078,216	135
8	116,942,821	128
9	108,145,351	116
10	106,383,598	118
11	110,171,769	130
12	85,358,539	109
13	84,419,198	105
14	81,345,643	103
15	84,633,453	109
16	77,906,053	94
17	76,506,943	95
18	66,141,439	84
19	65,312,493	109
20	75,796,353	82
21	69,173,390	83
22	61,848,140	88
23	53,376,148	80
24	65,020,233	78
25	44,060,403	68
26	51,750,746	79
27	48,749,334	67
28	46,084,206	61
29	51,998,940	69
30	88,516,663	146

the LD within a finite population over time. In total 2,000 generations were simulated using an effective population size of $N_{eff} = 100$, where the effective population size, consisting of $N_s = 50$ sires and $N_d = 50$ dams, was kept constant over generations and random mating was applied. The latter includes that each animal was allowed to mate twice. This was realized for each tested population genetic model (see below).
The genome of an offspring is typically created based on the genomes of the offspring's parents, whereby each parent has a maternal and a paternal strand inherited from its parents. The underlying mechanism are briefly described in the following. For each parent the following steps were applied to obtain a maternal or paternal strand for an offspring:

1. For each chromosome it was chosen with equal chance to start on the maternal or paternal strand.

2. The recombination rates between adjacent SNPs were calculated based on the genetic map.

3. To realize recombination events, the calculated recombination rates were compared with random values x drawn between zero and one. The following two cases were possible:
   $x < \theta$: a crossing over occurred and thus the strand changed from the paternal to the maternal, or vice versa.
   $x \geq \theta$: no crossing-over took place and thus no strand change occurred.

First, a population genetic model was applied without mutation and selection, which means that only genetic drift and recombination had an influence on the extent of the LD within the population over time. Second, a population genetic model was applied with mutation and without selection. It is common in such studies to disregard the influence of selection for simplification (e.g., Meuwissen et al., 2001).

Settings for population genetic model without mutation (drift model): To build the founder generation the SNP alleles were drawn by chance and the allele frequencies were $p = q = 0.5$.

Settings for population genetic model with mutation (mutation-drift model): In the founder generation all SNP alleles were set to zero, i.e., starting with a homozygous founder generation. During the simulation of generations the alleles get the chance to mutate to one, whereby two different scenarios were investigated:

- Mutation scenario 1: Each SNP has the chance to mutate once per generation, but no back-mutation is allowed, i.e., 0 to 1, but not 1 to 0.

- Mutation scenario 2: Each SNP has the chance to mutate once per generation, wherein, when a SNP position is drawn again in the following generations then it is allowed to mutate back, i.e., 0 to 1 and 1 to 0

Different mutation rates $m \in \{0.0025,0.00125,0.001,0.00025\}$ were applied for both scenarios, whereby $m = 0.0025$ and $m = 0.00025$ were chosen similar to Meuwissen et al. (2001).

For all settings of the population genetic models, the LD was calculated following Eq. 1.7 on page 7. After each generation the LD was recorded to study its behavior over time. Each population genetic model was replicated ten times for each setting. For this simulation study and for the simulation of genetic values and phenotypes in the next section different Fortran-77 programs were developed.

3.2.2 The alternative (SBML) approach for simulation

In this section, the realization of the SBML approach as well as the analysis set-up for the comparison between the conventional approach and the SBML approach is described. The following sections are adopted from Melzer et al. (2013b).

3.2.2.1 Population genetic model: mutation-drift model

Four hundred generations of a mutation-drift model (cf. mutation scenario 1, see above) with a constant effective population size of $N_{eff} = 100$ ($N_s = 50$, $N_d = 50$) were simulated employing random mating, whereby the mutation rate was set to $m = 0.0025$. Following the 400 initial generations, four additional generations were simulated without mutation and the population size was increased from 100 to 1,000 animals, which is common in the field of GS (e.g., Meuwissen et al., 2001). Here, a 50 half-sib mating design was applied (one sire mated with 20 dams), which corresponds to a mating design as it can be observed in the Holstein population (cf. Figure 1.1 on page 11). Generations 401 and 402 were used as training set (first offspring generation), generations 403 and 404 as test set (second offspring generation).

3.2.2.2 Simulation and analysis set-up for simulation approaches

The following simulation steps were applied. The number of QTL (n_{QTL}) was determined based on the used metabolome level model (Holzhütter, 2004). It is an erythrocyte metabolism non-linear ODE system model for human, which includes the glycolysis and pentose phosphate pathway (cf. Section 1.6.4.2 on page 24). The presence of all involved enzymes in *Bos taurus* was verified using the databases KEGG cow (Kanehisa and Goto, 2000) and Ensembl cow (Ensembl, 2008). While all 38 enzymes of this metabolome level were simulated numerically, only 23 enzymes, which cover parts of the glycolysis, gluthathione and pentose phosphate pathways in *Bos taurus*, were selected

to be influenced by 23 QTL. In addition, to work with larger numbers of QTL, we used the 10-fold quantity of QTL ($n_{QTL} = 230$, see details below). QTL positions were chosen randomly from all simulated SNPs with an MAF of at least 0.02 in generation 400. Furthermore, a reduced SNP data set was created from the complete SNP data set ($n_{SNP} = 52{,}273$), where every 10th SNP was included, but QTL positions were retained ($n_{SNP} = 5{,}227$). Combining the different numbers of SNPs and QTL resulted in four simulation scenarios:

Scenario 1: $n_{QTL} = 23$ and $n_{SNP} = 5{,}227$,

Scenario 2: $n_{QTL} = 230$ and $n_{SNP} = 5{,}227$,

Scenario 3: $n_{QTL} = 23$ and $n_{SNP} = 52{,}273$,

Scenario 4: $n_{QTL} = 230$ and $n_{SNP} = 52{,}273$.

Phenotypes were simulated based on different choices for broad-sense heritability $H^2 \in \{0.1,0.3,0.5\}$. For each scenario and heritability the set-up was replicated 100 times. The prediction precision, ρ, is defined as the correlation between simulated genetic values, g, from the test set, and predicted genetic values, $\hat{g}$. The following equation shows the prediction precision more formally:

$$\rho = cor(\hat{g},g). \tag{3.2}$$

We also investigated the impact of all 23 QTL on each metabolic outcome via regression analysis. In addition, the goodness of fit for this model was evaluated for all training data sets, scenarios and heritabilities, where the correlation between fitted values and residuals was determined using the function cor.test (Pearson's correlation coefficient) in R.

Simulating genetic value and phenotype - conventional approach: Following the conventional approach, the phenotype (y_i^{conv}) for an animal is simulated as:

$$y_i^{conv} = \sum_{j=1}^{n_{QTL}} (X_{ij}a_j + D_{ij}d_j) + e_i, \tag{3.3}$$

where $i \in \{1,...,n\}$ is the animal index. X_{ij} represents the design matrix for the additive effect a_j (allele substitutions effect), and D_{ij} is the design matrix for the dominance effect d_j. Entries in the design matrices depend on the observed marker genotypes at QTL j:

$$X_{ij} = \begin{cases} -1 & \text{genotype 11} \\ 0 & \text{genotype 12} \\ 1 & \text{genotype 22} \end{cases} \qquad D_{ij} = \begin{cases} 0 & \text{genotype 11} \\ 1 & \text{genotype 12,} \\ 0 & \text{genotype 22} \end{cases} \tag{3.4}$$

whereby $X_{ij} = 1$ means homozygous for the mutated alleles. The simulated additive genetic effect was drawn from a gamma distribution with shape parameter $\alpha = 0.42$ and scale parameter $\beta = 2.619$ in case of 23 QTL, and $\beta = 8.282$ in case of 230 QTL, following Meuwissen et al. (2001). The sign of the additive genetic effect was randomly drawn with equal chance. The dominance genetic effect was calculated as product of the additive genetic effect and the degree of dominance, which was drawn from a normal distribution with mean $m^* = 0.193$ and variance $\tau^2 = 0.097$ (Bennewitz and Meuwissen, 2010). The genetic value for an animal was composed as the sum of locus-specific genotypic effects as given in Eq. 3.3. Furthermore, genetic values of the training set and for the test set were separately standardized to obtain a simulated genetic variance $\sigma_g^2 = 1$. The phenotype for an animal was obtained by adding an error e_i to the genetic value. The error was drawn from a normal distribution $N(0,\sigma_e^2)$, for which the variance was determined according to the chosen $H^2 \in \{0.1,0.3,0.5\}$. In more detail, the residual variance σ_e^2 can be determined as follows (cf. Eq. 1.5 on page 5):

$$\sigma_e^2 = \frac{\sigma_g^2}{H^2} - \sigma_g^2. \tag{3.5}$$

For example, assume $H^2 = 0.3$ and $\sigma_g^2 = 1$, then the obtained residual variance is $\sigma_e^2 = 2.33$.

Simulating genetic value and phenotype - SBML approach: For our alternative SBML approach, we simulated a metabolome level between genotype and phenotype. A schematic representation of the conventional approach (A) and the SBML approach (B) is presented in Figure 3.1. The transition from the genotype to the metabolome level was realized as follows: a QTL influences a specific kinetic parameter k_{ij}, in our case mostly the V_{max} value of a specific enzyme. This means that the kinetic parameter changes depending on the genotype of the QTL coded in X_{ij} (cf. Eq. 3.4). In detail, $k_{ij} \in \{\Psi - 50\%,\Psi,\Psi + 50\%\}$, following Holzhütter (2004), correspondes to $X_{ij} \in \{-1,0,1\}$ if the sign of the additive genetic effect was positive in the conventional approach. Otherwise, the order of the values of the kinetic parameter was reversed, i.e., $k_{ij} \in \{\Psi + 50\%, \Psi, \Psi - 50\%\}$. Ψ was the default value of the kinetic parameter in the originally parameterized SBML model. The enzyme kinetics, and corresponding metabolites (in total 27 metabolites) that were affected are listed in Table 3.2. All other parameters remained unaffected in the SBML model.

The SBML model was implemented as a numeric simulation of the ODE system for the respective kinetic parameter settings using Matlab R2009b and the Matlab toolbox SimBiology R2009b (MATLAB, 2009). The SBML model was simulated until the metabolite concentrations reached the steady-state. After test runs, the maximum number of iterations (time) was set to 500. On the basis of the standardized equilibrium

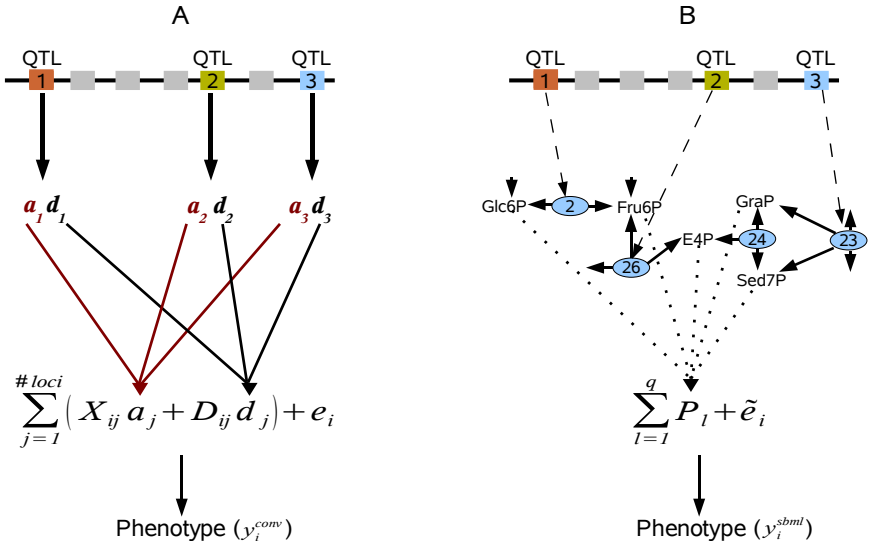


Figure 3.1: Schematic representation of the conventional approach (A) and the SBML approach (B).

metabolite concentrations, we simulated the phenotype (y_i^{sbml}) for an animal as:

$$y_i^{sbml} = \sum_{l=1}^{q} (P_{il}) + \tilde{e}_i. \tag{3.6}$$

P_{il} depicts the matrix of equilibrium metabolite concentrations, where $i \in \{1,...,n\}$ is the animal index, and $l \in \{1,...,q\}$ denotes the index for equilibrium metabolite concentrations belonging to specific enzymes influenced by simulated QTL. All equilibrium metabolite concentrations belonging to those metabolites (cf. Table 3.2) catalyzed by a specific enzyme were summed up, resulting in the specific metabolic outcome for simulated QTL. Further, q represents the total number of corresponding equilibrium metabolite concentrations; in some cases two metabolite concentrations were influenced by one enzyme. Furthermore, in some cases more than one column of P belongs to the same metabolite, i.e., the metabolite is catalyzed by more than one of the investigated enzymes. In our arbitrary mapping, the sum over all equilibrium metabolite concentrations, i.e., the sum of all 23 metabolic outcomes, results in the genetic value for an animal. Note that for this simulation approach of a GP map, the step from genotype to the metabolite concentrations is non-additive, whereas from the metabolite concentrations to the genetic value a purely additive step is implemented. Similar to the conventional approach, the genetic values for the training set and the test set were standardized separately. The phenotype was obtained by adding an error $\tilde{e}_i$, which was drawn from a normal distribution with mean zero and residual variance $\sigma^2_{\tilde{e}}$. The residual variance was again determined according to the chosen $H^2 \in \{0.1,0.3,0.5\}$.

Two sizes of SBML models were implemented, a 23-QTL model and a 230-QTL model. For

Table 3.2: The enzyme characteristics which were changed in the used SBML model. All other parameters were unaffected in this model. In addition, the corresponding number of the influenced metabolites based on the position of the obtained output of this model.

EC nr	Enzyme characteristic*	$\Psi - 50\%$	Ψ^*	$\Psi + 50\%$	Metabolite
EC 5.1.3.1	vRibPepi_Vmaxv21	2317	4634	6951	27
EC 2.7.1.40	vPK_Vmaxv12	285	570	855	2 + 17
EC 2.7.1.2	vHEX_Vmax1v1	7.9	15.8	23.7	3 + 4
EC 5.4.2.4	vBPGM_kDPGMv8	38000	76000	114000	14
EC 5.3.1.1	vTPI_Vmaxv5	2728.3	5456.6	8184.9	7
EC 1.2.1.12	vGAPDH_Vmaxv6	2150	4300	6450	11 + 12
EC 2.7.1.11	vPFK_Vmaxv3	119.5	239	358.5	4 + 6
EC 1.1.1.27	vLDHNADH_Vmaxv13	1400000	2800000	4200000	10 + 18
EC 2.2.1.1	vTrKet2_Vmaxv26	11.75	23.5	35.25	5 + 7
EC 2.7.6.1	vPPRPPS_Vmaxv25	0.55	1.1	1.65	31
EC 5.3.1.6	vRibPiso_Vmaxv22	365	730	1095	28
EC 2.7.4.3	vAK_Vmaxv16	690	1380	2070	4 + 22
EC 4.2.1.11	vENO_Vmaxv11	750	1500	2250	16
EC 5.3.1.9	vGPI_Vmaxv2	467.5	935	1402.5	5
EC 4.1.2.13	vALD_Vmaxv4	49.46	98.91	148.37	7 + 8
EC 3.1.3.13	vBPGP_Vmaxv9	0.27	0.53	0.8	9 + 13
EC 2.7.2.3	vPGK_Vmaxv7	2500	5000	7500	2 + 13
EC 3.6.1.5	vATPase_kATPasev15	0.84	1.68	2.52	4 + 9
EC 5.4.2.1	vPGM_Vmaxv10	1000	2000	3000	15
EC 1.8.1.7	vGSSGRD_Vmaxv19	45	90	135	20 + 26
EC 2.2.1.2	vTrAld_Vmaxv24	13.6	27.2	40.8	5 + 30
EC 2.2.1.1	vTrKet1_Vmaxv23	11.75	23.5	35.25	7 +29
EC 1.1.1.49	vG6PDH_Vmaxv17	81	162	243	19 + 23

Ψ - default value of the enzyme kinetic parameter;
* enzyme kinetic parameter values were adopted from Holzhütter (2004)

the 230-QTL model, the original model was replicated 10 times, yielding 230 independent QTL. For each replicate, the 23 enzymes available in cows were simulated as QTL as outlined above.

3.2.2.3 Predicting genetic values using fastBayesB

Prediction of genetic values was based on using the genotypes and phenotypes from the training set to estimate the genetic effect sizes. These genetic effect sizes were combined with the genotype from the test set to estimate the genetic value. We considered the fastBayesB method an appropriate choice for our studies (Meuwissen et al., 2009), which is an iterative fast Bayesian approach to estimate additive genetic effects. An extended version of this method including non-additive genetic effects is described in Wittenburg et al. (2011). We implemented a fastBayesB analysis considering additive and dominance

genetic effects for simulated animals $i = 1,...,n$. The mixed model can be expressed as follows:

$$y = Xa + Dd + \epsilon, \tag{3.7}$$

with

y = vector of phenotypes $[y = (y_1,...,y_n)']$,
$X = (n \text{ x } n_{SNP})$-design matrix of the additive genetic effects $a = (a_1,...,a_{n_{SNP}})'$,
$D = (n \text{ x } n_{SNP})$-design matrix of the dominance genetic effects $d = (d_1,...,d_{n_{SNP}})'$,
ϵ = residuals.

The residuals were assumed to be independently and normally distributed $\epsilon_i \sim N(0,\sigma^2_\epsilon)$. Entries of the design matrices are random variables and depends on the observed SNP-genotypes. SNP-genotypes are coded as presented in Eq. 3.4 on page 50, where the homozygous with the more frequent allele is coded as one. It is also assumed that there is LE between the SNPs and that genetic effects at different loci are independently distributed. The additive and dominance genetic effects were re-parameterized as follows (notation was adopted from Wittenburg et al., 2011):

$$\begin{aligned} Xa &\rightarrow \tilde{X}_a g_a, \\ Dd &\rightarrow \tilde{X}_d g_d, \end{aligned} \tag{3.8}$$

to prevent the estimation of covariances between them. For this, we applied the orthogonal decomposition of the genetic values g_s, $s \in \{a,d\}$, the method of Álvarez Castro and Carlborg (2007), according to Wittenburg et al. (2011). The fastBayesB algorithm involves prior assumptions for genetic effects. The prior distribution of a genetic effect $g_{s,j}$ at locus $j \in \{1,...,n_{SNP}\}$ is a mixture of the double exponential distribution (i.e., Laplace distribution) with zero expectation and the point mass at zero. The probability of having a zero genetic effect at some locus is $g_{s,j} = 1 - \gamma$. Hence, γ represents the proportion of QTL to SNPs, and the algorithm requires a specification of this parameter. Another parameter, which can be set by the user, is λ. This parameter mirrored the prior uncertainty of a genetic effect and was fixed ($\lambda = \sqrt{2 \cdot n_{SNP} \cdot \gamma}$).
As the true number of QTL for a given trait is generally unknown, the following set of plausible values for $\gamma \in \{0.1,0.05,0.025,0.01,0.005,0.001,10^{-4},10^{-5}\}$ was tested for each run of fastBayesB. The resulting variation of prediction precision in the sets was evaluated to mirror the sensitivity of the algorithm to different choices of γ. The optimal γ was determined over the corresponding replicates, resulting in largest mean prediction precisions (cf. Eq. 3.2). The genetic variance σ^2_g was determined as additive genetic variance σ^2_a plus dominance genetic variance σ^2_d.
The maximum number of fastBayesB iterations was set to 1,000. Also, SNP alleles with MAF < 0.01 were excluded from the analysis, but SNP alleles that were not in HWE were kept.

3.2.3 The metabolite approach for prediction

In this section, the workflow of the metabolite approach is presented. The approach was based on assigning weights to the SNPs according to importance of the metabolites to their impact on the phenotype. The most suitable measure for weighting SNPs is also shown. The analysis set-up also allows us to study the behavior of the prediction if only a part of the metabolome is measured, especially regarding our experimental data.

3.2.3.1 Data sets

For this study, the data sets were simulated based on the SBML approach as described in Section 3.2.2.2, whereby the following settings were chosen: $n_{SNP} = 52{,}273$, $n_{QTL} = 230$ and $H^2 \in \{0.1{,}0.3{,}0.5\}$. Here, the 230-QTL model was used, which has an influence on 230 enzymes and catalyzes 270 metabolites in total. In this case, the equilibrium levels of these metabolites were used without standardization to simulate a genetic value (cf. Section 3.4.3 for more information) and no additional noise was added to the simulated metabolite profiles.

3.2.3.2 Workflow

To achieve a weighting scheme for predicting the genetic values, three steps were necessary as shown in Figure 3.2. The first step is to use the metabolite profiles to predict the phenotype. We used a random forest (RF) regression method, which contains a technique of variable selection, i.e., we applied Random Jungle (version 0.7.2), a fast C-Implementation by Schwarz et al. (2010), using default settings. The variable importance is a measure that quantifies how much of the investigated phenotype is explained from each metabolite. The obtained variable importances (based on the Gini-index) of the metabolites were used for the next step.
Second, in the simulation case, we knew which enzymes catalyze the reactions involving the important metabolites. At this point, it was possible that one or two metabolites were catalyzed by the same enzyme. If two metabolites were catalyzed from the same enzyme, then the maximum value of the observed variable importances was used. For simulated SNPs we used the "real" positions on the genome and created a vector of weights for 52,273 SNPs, where the obtained variable importances were assigned to the QTL positions. The latter was possible because we knew which QTL had affected which enzyme. All other SNPs were weighted with the neutral weight value of one.
In the third step, we estimated the genetic effect sizes for each SNP using fastBayesB including additive and dominance genetic effects (as described in Section 3.2.2.3), whereby γ was set to 0.0001. The vector of weights was incorporated in the style of a weighted regression approach. The last step was also applied for the SBML approach without weighting the specific QTL, which resulted in a classical analysis and thus served as reference value.

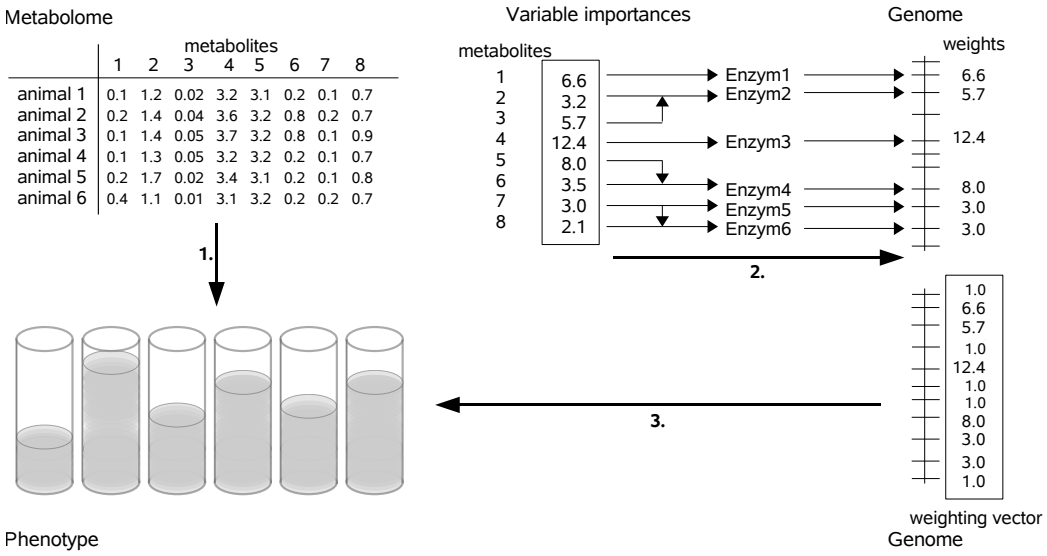


Figure 3.2: Schematic representation of the weighting approach.

3.2.3.3 Weighting approaches

Four different scenarios were evaluated to find an appropriate setting for weighting SNPs as well as to investigate the importance of the measured part of the metabolome.

Weighting scenario 1: All 230 QTL positions were weighted with a constant weight, and all other SNPs were neutrally weighted with 1. Weights (w) were elements of $w \in \{1, 2, \ldots, 10\}$. This scenario was mainly used to determine a suitable weight range to enable a transformation of the observed variable importances for the following scenarios.

Weighting scenario 2: The 270 simulated metabolites were considered "measured" and termed identifiable metabolites. In this scenario all metabolites were used to predict the phenotype. The resulting variable importances from RF were used as weights for the known 230 QTL; all other SNPs got the neutral weight.
Investigations of the observed variable importances of this scenario revealed that metabolite 14 (cf. Table 3.2) had the highest impact on the investigated phenotype for all used heritabilities and over all replicates. In addition, metabolite 14 was included ten times and at least one of these occurrences had the highest variable importance.

Weighting scenario 3: Only 20% of the 270 identifiable metabolites were selected randomly. In this scenario, metabolite 14 was always included. The other 80% of the metabolites were non-identifiable and were simulated at random.

Weighting scenario 4: Analogous to scenario three, whereby metabolite 14 was always excluded.

For each weighting scenario, including the case without weights for QTL, and broad-sense heritability, the set-up was replicated ten times.

3.3 Results for simulation studies

3.3.1 Analysis of the simulation approaches regarding a suitable LD

In this section, suitable levels of LD for a genome simulated using the experimental SNP chip annotation are presented. Different settings were applied for the tested population genetic models (cf. Section 3.2.1.2 on page 47). Figure 3.3 shows the results for the population genetic models using different settings, whereby the same mutation rate ($m = 0.0025$) was used for both scenarios with mutation. This Figure shows clearly that the population genetic model without mutation produces a considerably larger LD compared to both other scenarios where mutation was applied. This holds also true for other investigated mutation rates (cf. Figure 3.4). The observed curve for mutation scenario 1 shows similar behavior to the obtained curve without mutation. Neither case results in an increase-decrease equilibrium. In comparison, mutation scenario 2 with mutation resulted in an increase-decrease equilibrium, which means an equilibrium between genetic drift, mutation, and recombination. The highest observed mean value of LD was $r^2 = 0.201$ without mutation, $r^2 = 0.119$ for mutation scenario 1 ($m \in \{0.0025, 0.00125\}$) and $r^2 = 0.105$ for mutation scenario 2 ($m = 0.00125$). The mean LD values for the different

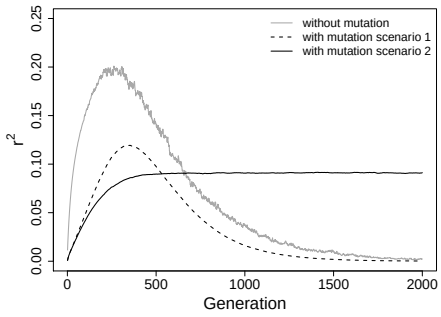


Figure 3.3: The mean LD values (r^2) over 2,000 generations of ten replicates for the settings without mutation and for both scenarios with mutation using the mutation rate $m = 0.0025$ are presented.

mutation rates for mutation scenario 1 in Figure 3.4 A and for mutation scenario 2 in Figure 3.4 B are presented. In both Figures can be seen that the chosen mutation rate has an impact on the extent of the LD.

3.3.2 Conventional versus SBML simulation approaches

In this section the analysis results are presented for the comparison of data simulated with the conventional and SBML approach, where a more complex GP map was used.

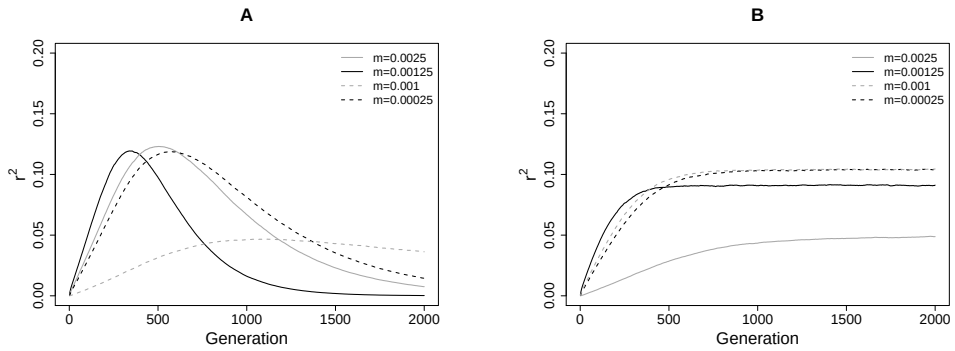


Figure 3.4: The mean LD values (r^2) over 2,000 generations of the ten replicates for mutation scenario 1 (A) and mutation scenario 2 (B) using different mutation rates.

The presented analyses based on the population genetic model with mutation ($m = 0.0025$) resulted in a mutation-drift model. This was chosen although it does not deliver the highest LD and does not correspond well to the desired LD of 0.2, but this scenario is more realistic than the drift model without mutation.

A comparison of observed LD between adjacent SNPs over the whole genome for the experimental and simulated SNP-genotypes is presented in Figure 3.5. In this Figure the observed LD for the experimental data set (including 43,079 SNPs; cf. Section 2.2.1 on page 31) and an example set of the simulated data after filtering the SNPs (around 46,500 SNPs) is shown. Comparative investigations of simulations regarding LD in training sets, excluding SNPs with MAF less than 1% (on average 5,688 SNPs), showed an average $r^2 = 0.14$. The average LD in the test sets was $r^2 = 0.15$ after discarding SNPs with MAF less than 1% (in average 5,826 SNPs). Additionally, in training sets as well as in test sets, only one SNP was not in HWE on average. In comparison, an LD of $r^2 = 0.21$ was obtained in our experimental data.

To obtain an optimal choice for the parameter γ, which was required for the fastBayesB estimation algorithm, different γ-values were implemented to analyze the four simulation scenarios and for three values of H^2. In general, it was observed that not every γ-value is appropriate for each scenario and heritability in the conventional approach and the SBML approach. For extreme choices of data or parameters, the fastBayesB algorithm aborts; for example, for $n_{SNP} = 52{,}273$ and $n_{QTL} = 23$, each value of H^2 and $\gamma = 0.1$, more than 74% of the 100 replicate runs aborted for both simulation approaches. A detailed list is presented in Appendix A.1 on page 141.

In Table 3.3, prediction precisions, simulated and estimated variance components, and corresponding standard deviations are listed for all tested scenarios and heritabilities regarding the optimal γ-value for both approaches. In general, $n_{SNP} = 5{,}227$ showed a larger prediction precision than $n_{SNP} = 52{,}273$. In addition, $n_{QTL} = 23$ showed a larger prediction precision than $n_{QTL} = 230$. The quantity of QTL had more influence

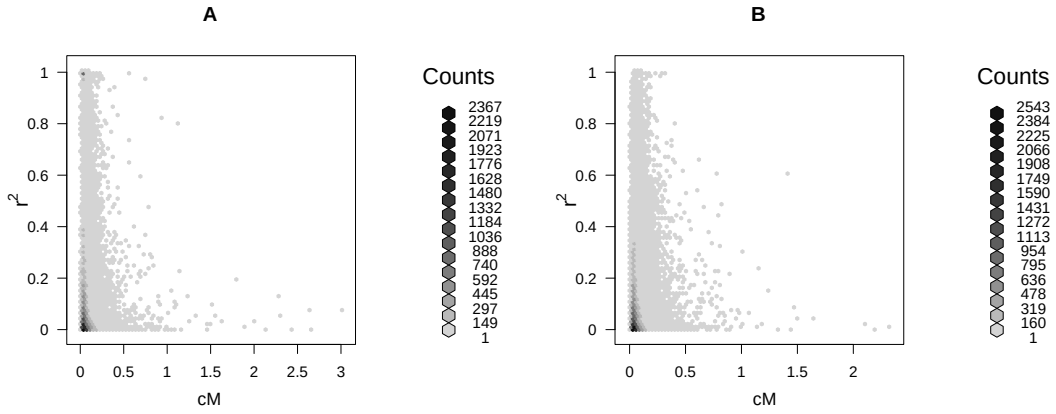


Figure 3.5: Comparison of linkage disequilibria between experimental and simulated data. The graphic shows the distribution of LD as correlations between adjacent SNPs for (A) experimental data set; (B) example simulated data set.

on the prediction precision than the quantity of SNPs. In more detail, in all investigated scenarios it was observed that the mean prediction precision was at least 3.75% lower for the SBML approach compared to the conventional approach. Estimated genetic variance components approached the true values for increasing values of simulated heritability. The estimated proportions of additive genetic variance to total genetic variance, σ_a^2/σ_g^2, were high compared to the proportion of dominance to total genetic variance. The estimated additive genetic variance σ_a^2 can be used to evaluate the degree of linearity of both simulation approaches, which is at least 5.88% lower for the SBML approach compared to the conventional approach for all investigated scenarios. Figures 3.6 A-B show the simulated and estimated additive and dominance genetic effects for an example data set using the conventional approach based on $H^2 = 0.3$, $n_{QTL} = 23$ and $n_{SNP} = 52{,}273$. It was observed that large simulated genetic effects were better detected than small genetic effects by the fastBayesB method. In comparison, Figures 3.6 C-D show the estimated additive and dominance genetic effects for the comparable SBML approach. Here, sizes of the simulated genetic effects were unknown. Hence, in an additional analysis involving only the 23 simulated QTL and genetic values of the simulated trait, we obtained estimates for the implicitly simulated genetic effects for the SBML approach. To characterize possible deviations from linearity in the SBML approach, we estimated the genetic effect sizes of all 23 simulated QTL on the observed metabolic outcome of each single QTL influenced enzymatic reaction. For an example data set, which was also the basis for Figures 3.6 C-D, the impact of all QTL on different metabolic outcomes is presented in Figure 3.7. Analyzing all 100 data sets, two different kinds of GP mapping were observed on the level on metabolite concentrations:

- First, the QTL had no clear impact on the metabolic outcome, whereas other QTL positions did. As an example, QTL1 had no specific impact on the corresponding

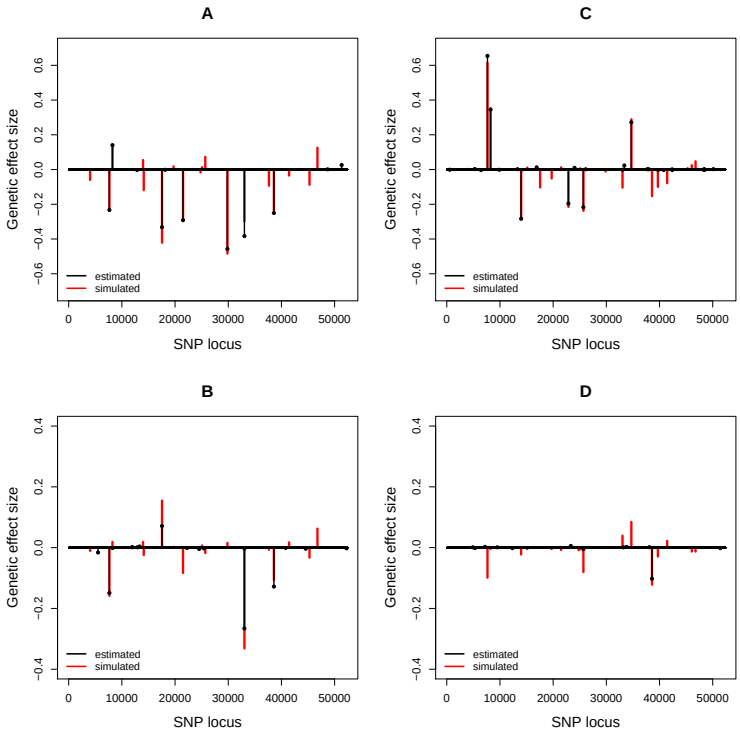


Figure 3.6: The estimated main genetic effect sizes for the conventional approach (left) and the SBML approach (right). All figures are based on an example data set with $n_{SNP} = 52{,}273$, $n_{QTL} = 23$, $H^2 = 0.3$ and the optimal γ-value. Estimated additive genetic effects (A) and dominance genetic effects (B) in the conventional approach. A filled circle was plotted for each genetic effect $> 10^{-4}$. In comparison, estimated additive genetic effects (C) and dominance genetic effects (D) in the SBML approach. Here, the implicitly simulated main genetic effect sizes were estimated using the 23 QTL to predict the corresponding genetic values. The observed estimated genetic effect sizes were plotted in red.

metabolic outcome, whereas, for example, QTL18, QTL22, clearly had an impact on the metabolic outcome belonging to the enzymatic reaction parameterized by QTL1.

- Second, the QTL had a clear impact on the metabolic outcome as well as other QTL positions; for example, this is the case for QTL18 and QTL23.

For all 100 training data sets, all scenarios, heritabilities and for the corresponding optimal γ-values, we investigated how good the linear model fitted the simulated data for both simulation approaches. Results are presented in the Appendix A.2 on page 142. We observed, except in one case, that the linear model fitted all simulated data sets similarly, and no significant difference in the simulation approaches was found.

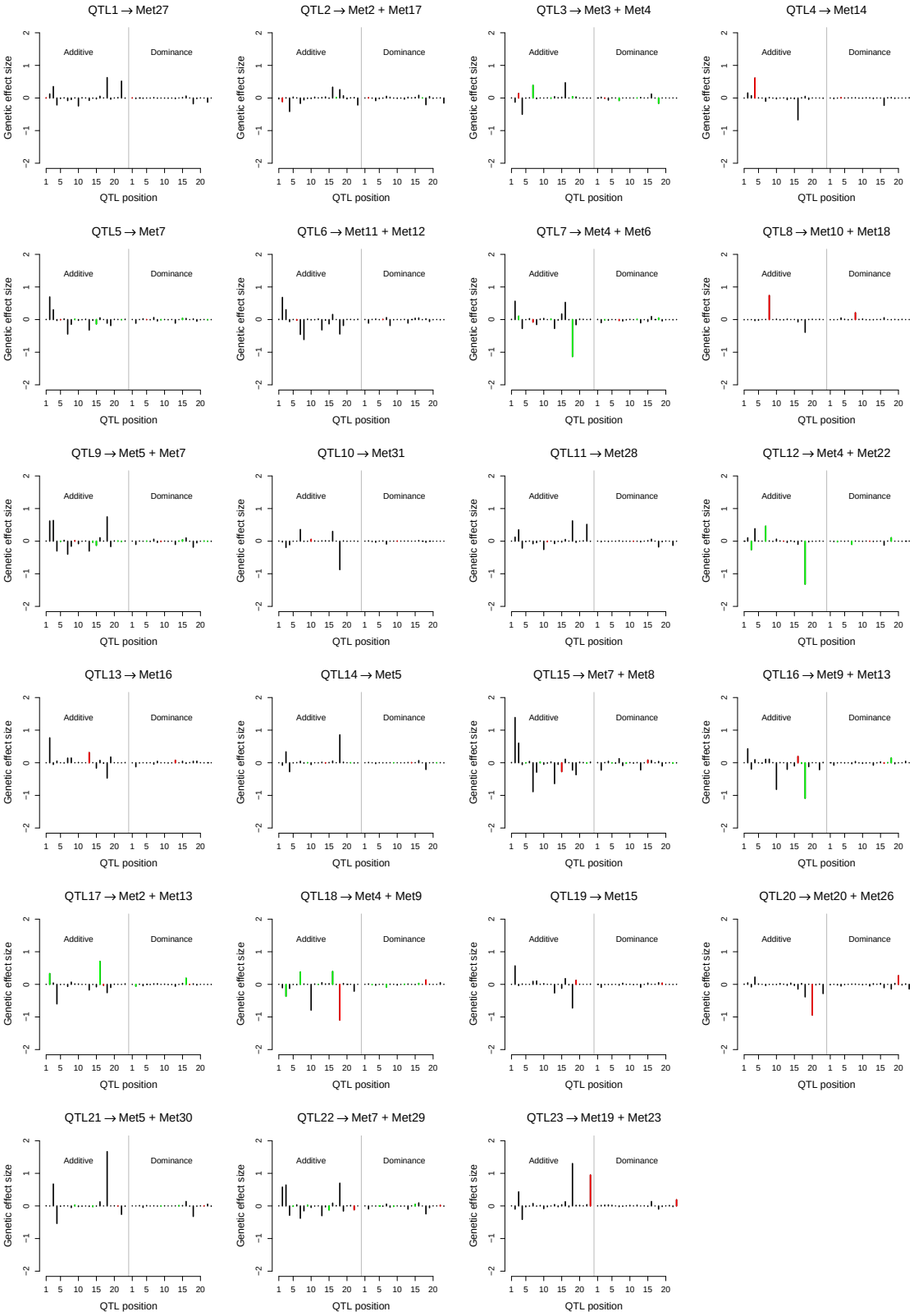


Figure 3.7: The estimated main genetic effect sizes for all QTL for each metabolic outcome. Each QTL is numbered and its specific metabolic outcome is presented. The metabolic outcome is split into the participating metabolites (Met). QTL positions which share the same metabolite for their belonging enzymes are marked in green and the corresponding QTL position is marked in red.

Table 3.3: For the optimal γ-value, the average estimated variance components and in parentheses the corresponding standard deviations are given for 100 replicates and compared with the simulated variance components (italic) for the different scenarios.

Approach	n_{QTL}	n_{SNP}	H^2	σ_g^2	σ_a^2	σ_d^2	σ_e^2	$\hat{\sigma}_g^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_d^2$	$\hat{\sigma}_e^2$	$\hat{H}$	ρ
Conventional	23	5227	*0.1*	*1*	*0.95*	*0.05*	*9.00*	0.78 (0.17)	0.77 (0.17)	0.01 (0.03)	8.92 (0.28)	0.08	0.86 (0.06)
			0.3				*2.33*	0.93 (0.09)	0.89 (0.10)	0.03 (0.04)	2.33 (0.07)	0.29	0.96 (0.02)
			0.5				*1.00*	0.97 (0.07)	0.93 (0.08)	0.04 (0.04)	1.01 (0.03)	0.49	0.98 (0.01)
		52273	*0.1*				*9.00*	0.73 (0.18)	0.71 (0.18)	0.02 (0.03)	8.92 (0.29)	0.08	0.80 (0.10)
			0.3				*2.33*	0.90 (0.09)	0.87 (0.10)	0.03 (0.04)	2.34 (0.08)	0.28	0.94 (0.03)
			0.5				*1.00*	0.96 (0.07)	0.92 (0.08)	0.04 (0.04)	1.02 (0.04)	0.48	0.97 (0.01)
	230	5227	*0.1*	*1*	*0.94*	*0.06*	*9.00*	0.30 (0.11)	0.21 (0.09)	0.09 (0.03)	7.17 (0.29)	0.04	0.50 (0.10)
			0.3				*2.33*	0.71 (0.09)	0.68 (0.08)	0.03 (0.02)	2.32 (0.09)	0.23	0.75 (0.06)
			0.5				*1.00*	0.87 (0.08)	0.83 (0.08)	0.04 (0.03)	0.94 (0.04)	0.48	0.85 (0.03)
		52273	*0.1*				*9.00*	0.11 (0.08)	0.08 (0.07)	0.02 (0.02)	6.75 (0.30)	0.02	0.45 (0.11)
			0.3				*2.33*	0.59 (0.10)	0.58 (0.10)	0.01 (0.02)	2.59 (0.10)	0.19	0.66 (0.08)
			0.5				*1.00*	0.82 (0.08)	0.77 (0.08)	0.05 (0.03)	0.93 (0.04)	0.47	0.77 (0.05)
SBML	23	5227	*0.1*	*1*	-	-	*9.00*	0.73 (0.16)	0.70 (0.15)	0.03 (0.05)	9.04 (0.30)	0.07	0.82 (0.05)
			0.3				*2.33*	0.88 (0.08)	0.83 (0.08)	0.05 (0.03)	2.39 (0.07)	0.27	0.92 (0.02)
			0.5				*1.00*	0.93 (0.08)	0.86 (0.08)	0.06 (0.03)	1.00 (0.04)	0.48	0.94 (0.01)
		52273	*0.1*				*9.00*	0.68 (0.17)	0.66 (0.16)	0.02 (0.04)	9.05 (0.31)	0.07	0.77 (0.08)
			0.3				*2.33*	0.85 (0.09)	0.81 (0.09)	0.04 (0.03)	2.40 (0.08)	0.26	0.90 (0.03)
			0.5				*1.00*	0.90 (0.09)	0.85 (0.09)	0.05 (0.03)	1.07 (0.04)	0.46	0.93 (0.02)
	230	5227	*0.1*	*1*	-	-	*9.00*	0.22 (0.06)	0.13 (0.06)	0.08 (0.03)	7.30 (0.30)	0.03	0.37 (0.06)
			0.3				*2.33*	0.69 (0.09)	0.64 (0.09)	0.04 (0.02)	2.20 (0.07)	0.24	0.68 (0.04)
			0.5				*1.00*	0.76 (0.07)	0.73 (0.07)	0.02 (0.02)	1.10 (0.04)	0.41	0.78 (0.03)
		52273	*0.1*				*9.00*	0.04 (0.04)	0.02 (0.03)	0.02 (0.02)	6.87 (0.31)	0.01	0.32 (0.05)
			0.3				*2.33*	0.64 (0.09)	0.58 (0.08)	0.06 (0.04)	2.16 (0.08)	0.23	0.55 (0.07)
			0.5				*1.00*	0.73 (0.07)	0.69 (0.06)	0.04 (0.02)	1.00 (0.05)	0.42	0.69 (0.04)

3.3.3 Analysis of the metabolite approach for prediction

In this section the results are presented, whether an improvement can be obtained, when the whole or only a part of the metabolome is used for the genetic value prediction and which range of weights is suitable for weighting SNPs. Therefore different scenarios for weighting SNPs were proposed (cf. Section 3.2.3.3 on page 56). First, it was investigated which range of weight measures is suitable to obtain a high accuracy in the genetic value prediction. This investigation is based on weighting scenario 1. Hence, different weights were tested, whereby all known QTL were assigned the same weight measure. The observed mean prediction precisions for the different applied weight measures and H^2 are presented in Figure 3.8 A. This Figure shows that a weight measure of two and three are suitable weight measures to obtain a high mean prediction precision for all investigated H^2. In addition, the corresponding reference values for each heritability, i.e., QTL were neutrally weighted, are presented using the weight measure of one. In this context, the highest gain in prediction precision with respect to the observed prediction precision for the reference values was obtained for $H^2 = 0.1$. Based on the result of the

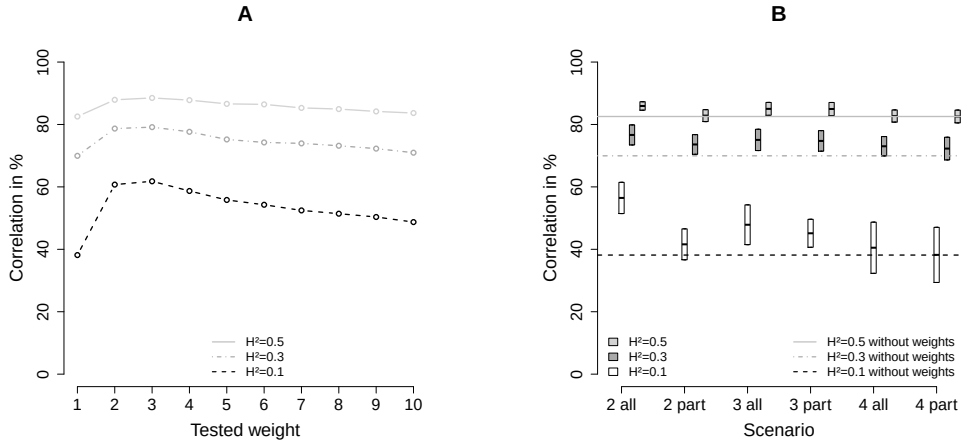


Figure 3.8: The observed precisions of genetic value prediction are shown using different weights. (A) The obtained mean prediction precisions for weighting scenario 1. (B) The observed prediction precisions (mean ± standard error) for weighting scenarios 2 to 4 depending on the proportion of identifiable metabolites: *all* means all identifiable metabolites of all 270 metabolites and *part* the best 10% of all 270 metabolites. In addition in weighting scenario 3 and 4 only the identifiable metabolites were used.

weighting scenario 1, the obtained variable importance values, viv, were scaled as follows for weighting scenario 2 to 4:

$$viv_{scale} = exp^{\frac{viv}{max(viv)}}, \tag{3.9}$$

where exp is the exponential function. After this transformation the weights lie in a range between 1 to 2.718. Two strategies were implemented to obtain a vector of weights

for weighting scenarios 2 to 4:

1. all identifiable metabolites were used; or

2. the top 10% of all metabolites were taken.

In addition, only the identifiable metabolites were further used in weighting scenarios 3 and 4. The observed prediction precisions are presented in Figure 3.8 B for weighting scenarios 2 to 4. In this Figure is obvious, that in weighting scenario 2, where all identifiable metabolites (2 all) were used, i.e., the whole metabolome was known, the best prediction precisions were obtained for all used H^2 as well as in comparison to the other tested scenarios in that case. Whereas, when only the best identifiable metabolites were selected in scenario 2 (2 part), the prediction precisions decreased significantly. Also, clear differences can be seen between weighting scenario 3 and 4 for both cases. Weighting scenario 3, which includes the known important metabolite 14, resulted in higher prediction precisions than weighting scenario 4 where metabolite 14 was completely excluded. The observed mean prediction precisions without using weights, i.e., reference values, are also presented in Figure 3.8 B (lines) for each H^2. We observed that in most tested weighting scenarios the prediction precisions are higher than for unweighted QTL.

3.4 Discussion

Methodological developments for algorithms in the field of GS are typically based on simulated data. In our contribution, as an alternative to the state of the art simplistic simulation approach, we investigated consequences of using a more complex, partly non-additive GP map, in comparison to a conventional GP map. In this context, the simulated genome was also created more realistically, as it was constructed based on the experimental SNP chip and thus investigations were necessary to obtain a more or less realistic LD. The presented comparisons between the conventional approach and the SBML approach revealed that the SBML approach produced lower prediction precisions and had a lower linear additivity compared to the conventional approach. The simulation of data using our SBML approach has an important advantage, as it allows further investigations that are not possible with the conventional approach. Using the metabolite approach on data simulated with the SBML approach, we observed that weighting QTL according to the importance of metabolites, which depend on their impact on the phenotype, can lead to an improvement in the genetic value prediction. The degree of improvement if only a part of the metabolome is considered for the genetic value prediction depends on the importance or relevance of the part for the investigated phenotype.

3.4.1 Simulation approaches to obtain an appropriate LD

The first aim of the simulation study was to obtain an appropriate LD to simulate a more realistic data set. For this setting we compared different parameters with those found in the literature. In general, when comparing linkage equilibria with and without mutation (Figure 3.3) we observed what we expected: a higher LD for simulated data without mutation. The reason is, that recombination and genetic drift are the only two influences on the LD in the drift model. In this model, we started with genetic variation in the founder generation. In the first generations, the genetic drift has a positive influence on LD, but over time the genetic variability gets lost, since favored alleles become fixed by the genetic drift. Recombination reduces LD and supports the genetic variation, but in this case the variation of alleles gets lost over time and no new genetic variation can be obtained by recombination. Recombination is weaker than genetic drift in the used settings, which is further favored by the small N_{eff} (cf. Section 1.2 on page 8). The latter holds also true for both mutation scenarios.
Compared to the scenario without mutation, the tested mutation scenarios started with no genetic variation within the population. The genetic variation was created in both mutation scenarios during the simulated generations. The behavior of LD over time in mutation scenario 1 (without allowing alleles to mutate back) showed similar behavior to the LD over time for the scenario without mutation. The added mutations in each generation lead to genetic variation, but the alleles become fixed over time and thus the genetic variation gets lost, because no new genetic variation can be obtained by mutation. The extent of LD in mutation scenario 1 depends on the used mutation rate (m), which also holds true for mutation scenario 2. In comparison to mutation scenario 1 in mutation scenario 2 the genetic variation is renewed by the fact that it is allowed to mutate back, which leads to an equilibrium of decrease and increase of LD over time. Summarizing, it was observed that small changes in simulation parameters (cf. Figure 3.4) had an enormous influence on the LD, its establishment, decay, maximum LD value, and development over time. Further, we were not able to reach the LD of 0.2 with a method of the mutation-drift model as found in literature. The reason for that may lie in the different ways to construct the genome. In the literature, typically chromosomes with equal length were simulated. In our study the genome was built in accordance to the bovine genome, which led to different sizes of chromosomes. Additionally, we used only 52,273 SNPs distributed over the bovine genome (about 30 M), whereas, for example, Calus et al. (2008) used 350,000 loci for a genome size of 3 M to get a high density of SNPs.
In conclusion, our experiences with mutation-drift simulations of genome-wide marker data showed that simulation parameters such as mutation rate, density of SNPs, and number of generations have to be chosen appropriately to result in LD values as found for experimental data.

3.4.2 Conventional approach versus SBML approach for simulation

In the conventional approach, the contributions of additive and dominance genetic effects were explicitly modeled and thus known. In contrast, for the SBML approach the influences of additive and non-additive genetic effects and their specific impacts on the total genetic variance were unknown and genetic effects were estimated based on the simulated genetic values.

Our comparison of fastBayesB results showed that the conventional and SBML approach were not similar regarding prediction precision and mostly show clear differences in estimated variance components (cf. Table 3.3). In general, however, the choice of heritability and simulated quantity of QTL and/or SNPs had a similar influence on the prediction precision for both simulation approaches. The prediction precision decreased with increasing quantity of SNPs, because the larger SNP set only included additional non-informative SNPs, without impact on the phenotypic variation. The estimated genetic effect of all these additional SNPs should be zero. The fastBayesB method estimated an effect size for each locus (iteratively) under the assumption of linkage equilibrium. Additionally, the LD (mean value of $r^2 = 0.15$ for neighboring SNPs) between our simulated SNPs was weak, such that we did not expect linkage influences on estimated genetic effects. Hence, estimation errors accumulated with an increasing number of SNPs. The quantity of simulated QTL had a major influence on the prediction precision, which is in agreement with the observation of Daetwyler et al. (2010) and Zhang et al. (2010). As the same amount of genetic variation in the simulation was now spread over more loci, most QTL had small effect sizes. Smaller genetic effect sizes were more difficult to detect by fastBayesB. The details of results depend on the value of H^2.

The SBML approach enables further research opportunities regarding the inner structure of the simulated GP map compared to the conventional approach. We found that some genetic effects were negligible if the sum was taken over all specific QTL outcomes (Figure 3.7). In our case, investigations of the specific QTL outcomes revealed two different mappings:

- The first type, involving QTL variation, showed no impact on the metabolic outcome of the enzymatic reaction parameterized by the QTL. This indicates that changes at the corresponding QTL position had no direct influence on this metabolic outcome. For example, QTL1 position appears to have a negligible effect on the specific investigated metabolic outcome of all 23 investigated enzymatic reactions (Figure 3.7) consistently over all data sets. A genetic variation at QTL1, however, is not without importance for the main trait: If mutations or diseases affect either the metabolome network model or the weights for the summation of single metabolites to yield the phenotype, variation at QTL1 may become measurable.

- In the second type of observed mapping, the corresponding QTL position affects its specific metabolic outcome as well as that of other QTL positions. In this case,

some of the QTL positions interact.

Comparing estimated genetic effects for an example data set for the genetic value prediction based on the 23 QTL in the SBML approach (Figure 3.6 C-D) with those for the single metabolites (Figure 3.7) that are summed up to build these genetic values, it can be concluded that some genetic effect sizes, which exist on the metabolome level, are negligible with respect to genetic value.
For the conventional and SBML approach, the goodness of model fit was evaluated; the used linear model explained both simulated data sets similar in almost all cases. Hence, the observation that the simulated data of the SBML approach can be well analyzed with a classical linear model, including additive and dominance genetic effects, can be traced back to the arbitrary simple GP mapping from the metabolome level to the genetic value in the SBML approach. We conclude that for our chosen simulation approach, the SBML approach involves both a non-additive GP mapping as well as an additive part (metabolome to genetic value). In this context, we hypothesize that the genetic effects of the non-additive part possible lead to small deviations from a clear additive GP map for the phenotype. To decipher the details of these interwoven influences is certainly a rewarding field for future study.

3.4.2.1 The more realistic simulation approach

Our set-up of the SNP data sets was based on annotated SNP positions, and we used the actual lengths of the bovine chromosomes. This is different from most approaches recently chosen, where chromosomes have equal size, and mostly 3 to 10 chromosomes were simulated as discussed earlier. Our set-up generated a distribution of LD values for adjacent SNPs similar to the experimental data (Figure 3.5).
To simulate more realistic genetic values, several further opportunities exist (see below). We decided to integrate the level of the metabolome between genotype and phenotype and kept the construction of the genetic value as simple as possible: each QTL directly influences only one kinetic parameter per enzyme. However, there were a lot of indirect influences detectable (implicitly simulated biological epistasis). Further, taking the sum of equilibrium enzyme products is a simple strategy to simulate genetic values of a "complex trait", such as for example milk fat. The following alternative approaches might be conceived:

1. Genetic variation at a specific QTL may influence more than one enzyme parameter at a time. This would allow for concrete pleiotropy. Also explicit epistasis could be a possible extension, as proposed by Long et al. (2010) or Ober et al. (2011), but these authors employed statistical epistasis.

2. Genetic values could be directly constructed from multiple metabolite concentrations in various other ways, e.g., take all metabolic outcomes (use once), eliminate the second linear step or find a better transformation.

3. The most advanced possibility of simulating phenotypes would certainly be to implement a systems biology model including cell, organ, and physiology levels, which could lead to more realistic, implicit GP mappings (e.g., Nomura, 2010).

In addition, Pinna et al. (2011) proposed another alternative of simulating phenotypes as single gene expression values, embedded within the non-linear ODEs network similar to other gene expression values. This approach represents the other extreme of phenotype simulation, compared to our proposed SBML approach. Our choice of integrating over a larger number of metabolite concentrations could be interpreted as an approach to simulate complex phenotypes.

3.4.2.2 Sensitivity of fastBayesB

The parameter γ of the fastBayesB method often has a significant influence on the results of analyses, especially on the prediction precision. Therefore, different γ-values were tested to study the influence on the performance of the fastBayesB method with respect to different simulation approaches. The optimal γ-value was determined by using the γ-value with the largest prediction precision covering a certain set of γ-values. For the conventional and the SBML approach, it can be summarized that the optimal γ-value was mostly lower than the simulated proportion of QTL to SNPs. The range of γ-values, which was appropriate for $n_{SNP} = 5{,}227$, was in the interval $[0.0001; 0.05]$, and for $n_{SNP} = 52{,}273$ the range was $[0.00001; 0.001]$. In other cases, the fastBayesB algorithm did not converge or it aborted (cf. Appendix A.1). If the algorithm did not converge, the optimum was not reached within the 1,000 iteration steps. There are several possible reasons for abortion, that will not be discussed in this thesis because the fastBayesB was only applied and no methodological improvement or optimization was realized. In addition, possible reasons were discussed in Wittenburg et al. (2011).

3.4.3 The benefit of using the metabolite approach for prediction

In general, the results demonstrate the feasibility of an integrative bioinformatics approach to enhance genetic value prediction based on SNP data by incorporating information about the metabolome level, based on a weighting approach as a matter of principle. The choice of simulation parameters is, on the one hand, designed to meet experimental data structures (e.g., SNP distribution, number of metabolites). On the other hand, many simulated details of the GP map remain artificial choices (as mentioned above), and, hence, a proof of principle using experimental data is necessary and is realized in Chapter 4.

In contrast to the comparative study of the conventional approach and the SBML approach, here the obtained metabolite steady-state concentrations were not standardized before they were summed up to the genetic value. If metabolites are not standardized, the variance of each metabolite plays a role for the prediction of the investigated

phenotype using a regression method. In this context, in weighting scenario 2 it was observed that the variances of the metabolites differ strongly, whereby most metabolites show small variances except for metabolite 14. Similar behavior was observed for the replicates as well as for the different heritabilities in this weighting scenario. Otherwise, if metabolites are standardized, they have equal weights regarding their variances and thus the variable importance depends on the correlation between metabolites and the investigated phenotype. The latter is used in experimental data where it is of interest to find biological relevance, e.g., between metabolites and phenotypes or other OMICs relationships. However, in the presented simulation study, it was not of interest to find "biological relevance", but rather to investigate what happens if QTL are weighted according to the considered part of the metabolome, which depend on the impact of the corresponding metabolites on the phenotype. Hence, it was decided to use the approach without standardization of metabolites for all tested scenarios, because this enables us to include or exclude metabolites that have high variable importances for the investigated phenotype in a simple way for all replicates and heritabilities.
In weighting scenario 1, where QTL were weighted with equal weights we could observe that with increasing weight measures the prediction precision first increased to an optimum, before starting to decrease again (cf. Figure 3.8 A). These investigations were necessary to obtain a range of suitable weight measures. Further, clear differences in prediction precisions were observed depending on the part of the measured metabolome as expected. It was expected that the highest prediction precision is obtained for using the whole metabolome (weighting scenario 2), because all known information are used. It was also expected that if a part of the simulated metabolome is used that is not primarily important for the investigated phenotype (weighting scenario 4) then it is possible to see only small or no improvements regarding the prediction precision, whereas if the metabolome part contains important metabolites for the phenotype then an improvement can be expected (cf. Figure 3.8 B).
In general, the obtained results of our simulation study are encouraging, and we envision testing further possibilities of optimization, e.g., regarding the actual rescaling of the variable importances to yield an appropriate weighting vector, using experimental data.

3.5 Summary

Our experience simulating more realistic data with respect to the experimental data revealed that simulation parameters such as mutation rate, distribution of SNPs, size of chromosomes, and population genetic parameters (e.g., mutation-drift model, number of generations) have to be chosen appropriately to result in LD values as found for experimental data.
Furthermore, our alternative SBML approach was presented to simulate data based on a GP map, designed to be more realistic, including additional information on the

metabolome level and compared to data simulated using the conventional approach. Different scenarios were investigated: smaller prediction precisions (at least 3.75%) were observed for the SBML approach compared to the conventional approach. Also the degree of linearity (σ_a^2) was less (at least 5.88%) for the SBML approach compared to the conventional approach. To summarize, simulating a more complex GP map including a molecular level allows us to study the processing of variation from the genetic to the phenotype level in more detail and may prepare the basis for the development of modern methods of GS. Data simulated with the proposed SBML approach offers further investigation opportunities as exemplified by our proposed metabolite approach, and can be used for methodological development. Compared to the conventional approach, these additional possibilities make simulation approaches like the proposed SBML approach eligible for improving the genetic value prediction for experimental data. Furthermore, the non-additive genetic effects may be exploited by modern methods in the field of GS using this type of strategy.

Finally, the additional use of the simulated metabolome level revealed that it is possible to improve genetic value prediction, and that the degree of improvement mainly depends on the considered part of the metabolome. Based on these findings, we applied the metabolite approach in a similar manner for our experimental data in Chapter 4. Summarizing, such kinds of simulation studies could help to understand, to interpret or to estimate the extent of an improvement that can be possibly expected, especially in view of the experimental data, where all three kinds of system-levels are available (genotype, molecular level, phenotype).

4 Analyses of experimental data

This chapter contains all investigations regarding the experimental data and is consistently structured in three parts.
The first part focuses on the purely conceptual comparison between the experimental data set and two simulated data sets (conventional approach and SBML approach) from Chapter 3 regarding analysis results obtained by fastBayesB. For this comparison, three milk traits were chosen. Parts of this chapter has been published in Melzer et al. (2013b).
The second part focuses on different relations between milk metabolites and milk traits as well as within each level. To enable a deeper understanding of these relations various statistical analysis (uni- and multivariate) methods were applied. In particular, the relations between milk metabolites and milk traits were of interest to find sets of metabolites eligible to predict the investigated milk traits. The latter was also realized in order to enable analysis of milk traits from a metabolic perspective and to shed light on a possible functional background for some of the detected associations. Such functionally important metabolites can serve as biomarker candidates. The identification of biomarkers also plays an important role in the field of dairy science, where it is of great interest to improve, for example, the detection and prevention of diseases. For this purpose two machine learning methods were applied. Our intensive investigations on both levels revealed new associations. Parts of this chapter have been published in Melzer et al. (2013a).
In the third part the metabolite approach is applied on three selected milk traits. The metabolite approach was used in a similar way as presented in Chapter 3. In contrast to Chapter 3 where important SNPs were weighted, here, SNPs are selected with a genetic impact on important metabolites (show a high importance for the milk trait prediction), resulting in a SNP subset which is used for the genetic value prediction. The metabolite approach was compared to the classical approach (all SNPs) and the reduced classical approach (selected SNPs), using a special invariable analysis design to enable comparability between SNP subsets for the genetic value prediction. We observed that the metabolite approach resulted in a more similar prediction precision to the classical approach as the reduced classical approach for our analyzed milk traits. Moreover, SNPs close to or within known QTL regions were determined resulting in a QTL-SNP subset. This QTL-SNP subset enabled us to determine if SNPs selected by the reduced classical approach and the metabolite approach were enriched in these genome regions. This represents a possible measure for the relevance of selected SNPs. The corresponding analysis revealed that more selected SNPs were located in these genome regions when the metabolite approach was used. This part has been published in Melzer et al. (2013c).

4.1 Introduction

In this section, first additional relevant background information regarding investigations of milk metabolites and milk traits in the field of dairy science are presented before we proceed to describe the current state of the art to select important SNPs in the field of GS belonging to MAS. This section relates mainly to the second and third part of this chapter. The main focus in this chapter is on part two and three, whereas in part one another perspective is proposed to compare experimental and simulated data regarding their composition of genetic effects detected by fastBayesB.

4.1.1 Background information for the analysis of the three system-levels

In the last years, metabolomics (e.g., Fiehn, 2002; Krastanov, 2010) have played an increasingly important role in several research fields, e.g., plant research (Weckwerth, 2003; Saito and Matsuda, 2010) or clinical research such as oncology (Spratlin et al., 2009), and received steadily more interest, also in dairy cattle research. In this context a Biomarker is defined as "A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Atkinson et al., 2001). In the field of dairy cattle science, biomarkers for diseases, e.g., mastitis and ketosis, or for the state of health and management remain to be found. In this context, an increased milk yield is assumed to cause undesirable side effects, such as an increase in health problems (e.g., Rauw et al., 1998). Diseases often reduce milk yield and also lead to additional costs, e.g., drug treatment or veterinarian costs (Ingvartsen et al., 2003).
The standard MPT, which is carried out regularly for each dairy cow in a monthly rhythm, is used to monitor the quality and quantity of specific milk traits (cf. Section 1.5.3 on page 18). Milk traits are also used as biomarkers for nutrition management and state of health. For instance, high values of SCC are a biomarker for mastitis (Dohoo and Meek, 1982; Schukken et al., 2003). However, the traditional milk traits used as biomarkers for the state of health are not sufficiently sensitive in view of diagnostic efficiency (e.g., SCC; Viguier et al., 2009). Even if for example acetone is an accepted biomarker for ketosis (Geishauser et al., 2000; Enjalbert et al., 2001), it is increased only if the disease is already acute. Klein et al. (2012) reported that no biomarker was available that shows long-term prognostic potential for ketosis, which will possibly hold true for other diseases. They proposed that the milk glycerolphosphocholine to phosphocholine ratio can be used to indicate a risk of ketosis. For mastitis, lactic acid was proposed as a potential biomarker (Farr et al., 2002; Davis et al., 2004). Also, it would be desirable to replace invasive diagnostics such as monitoring of rumen functions (Fievez et al., 2003; Vlaeminck et al., 2005) with noninvasive tests, such as a milk metabolome assay. Cabrita et al. (2003) found that levels of odd-chain fatty acids in milk have the potential to noninvasively monitor rumen function, and it was shown that heptadecanoic acid is a possible biomarker

candidate for protein deficiency in the feed. Prognostic markers or biomarker candidates are also sought for other diseases and management problems. Hence, it seems promising to find metabolites which can be used as biomarkers to improve diagnostic tools. However, to our knowledge, only few studies have been published regarding the association between milk metabolite profiles and milk traits obtained by MPT or the correlations among metabolites, e.g., Klein et al. (2010). Mostly, the correlation between single or groups of metabolites and single milk traits of interest were investigated in recent literature. In the field of dairy science, principal component analysis is often applied as a first unsupervised analysis approach (Sugimoto et al., 2012). In other fields, e.g., plant science, different multivariate analysis methods are typically used to explore the data. Sugimoto et al. (2012) give an overview of the current state of the art regarding such methods. To our knowledge analyses regarding multivariate correlations between sets of milk metabolites and milk traits from the MPT, as well as between metabolites, are lacking.

In general, a metabolite can be considered a new molecular milk trait, and genetic effects on it may be analyzed with estimation methods from the field of GS (e.g., Meuwissen et al., 2001; Goddard and Hayes, 2007). In a recent study, milk metabolites were considered new molecular traits and their genetic variability was investigated (Wittenburg et al., 2013).

Today it is common to genotype only elite animals, mostly bulls, because of the cost of a high-density panel e.g., Illumina® SNP Chip 777K. Hence, it is also of interest to design low-density SNP panels (3K - 6K), based on SNPs selected from the high density SNP panel, which can be used for a broader screening. A low density SNP panel should cover as many traits associated with breeding goals (cf. Section 1.3 on page 9) as possible in order to obtain an appropriate prediction precision for several traits (Vazquez et al., 2010). To determine an appropriate SNP subset from a high-density SNP panel, different strategies were proposed in recent literature. For instance, Habier et al. (2009) proposed to use equally spaced SNPs to obtain a SNP subset for several traits. Weigel et al. (2009) used Bayesian Lasso to find an optimal SNP subset for one trait, in which SNPs were ranked based on their genetic effects. A similar study is presented by Moser et al. (2010) who used ridge regression and partial least squares regression (PLS) to find an appropriate SNP subset for several production traits. In the study of VanRaden et al. (2009), GWAS was applied to detect important SNPs. In this context, Weller and Ron (2011) reviewed that more significant SNPs with a genetic impact on the investigated trait were detected using GWAS as with traditional designs (cf. Section 1.5.2 on page 17). This will probably also hold true for the above mentioned approaches, except for using equally spaced SNPs. These kinds of investigations belong to the field of MAS, because only a (filtered) SNP subset of all SNPs is applied for the genetic value prediction.

4.1.2 Analysis of the three system-levels

In the second part of this chapter, the intense investigations of the 190 milk metabolites and 14 milk traits of 1,305 Holstein Friesian cows are presented. Here, three additional ratios based on the 11 measured milk traits (cf. Section 2.2.2 on page 34) were additionally investigated to cover the status of EB, which is known to depend on the stage of lactation (cf. Section 1.5.3 on page 18). We investigated the impact of influencing factors on metabolite levels as well as on milk traits using univariate analysis methods. Especially the influencing factors of farm and day of lactation were analyzed in greater detail using multivariate analysis methods, since both influencing factors have an impact on the metabolic state of the cow (cf. Section 1.5.3). The main focus of our analyses was on the relations between milk metabolites and milk traits. These relations were analyzed taking a univariate analysis approach on the one hand, using the Pearson's correlation coefficient (e.g., Klein et al., 2010). On the other hand, different multivariate analysis methods, e.g., clustering and two machine learning methods, were applied (Sugimoto et al., 2012). Correlation structures within and between milk metabolites and milk traits are also reported and detailed results are presented in this thesis. The detected milk metabolites or groups of metabolites that have a significant impact on an investigated milk trait can serve as possible candidates for biomarkers or biosignatures. However, to propose concrete biomarker candidates, a suitable study would also have to include the traits of interest, for instance disease data. This kind of data were not part of this thesis. Instead, we used the obtained milk traits as surrogates for interesting health or management traits. In the corresponding discussion we present possible functional backgrounds for some of the associations found for specific important milk metabolites. In the third part, our proposed metabolite approach is applied on three selected milk traits and compared to three other approaches. The proposed approaches to select important SNP subsets for the genetic value prediction in recent literature (see above) is based on the classical GP map. In contrast, we propose that prediction precision may increase if SNP subsets determined for milk metabolites, which have a significant impact on the milk trait of interest, are used for the genetic value prediction. It is likely that an important milk metabolite is explained by a smaller number of QTL, i.e., has a less complex underlying genetic architecture, than a complex milk trait. Hence, it is possible that important SNPs, which have little genetic effect on a milk trait, may have stronger genetic effects on the corresponding important milk metabolites. It is expected that SNPs with a strong genetic effect on a milk trait (e.g., the SNP in the region of DGAT1 has a known large impact on fat; Weller and Ron, 2011; Grisart et al., 2004) also show a strong genetic effect on at least one of the determined important metabolites. To address these presumptions, our metabolite approach employs different sub-steps, in which the metabolome level was considered for the genetic value prediction in addition to the SNP information (similar to Chapter 3). In contrast to Chapter 3, where important SNPs were weighted, here, SNPs are selected with a genetic impact

on important metabolites, resulting in a SNP subset which is used for the genetic value prediction. Prediction precisions are compared for the following SNP subsets: metabolite SNPs, all SNPs (classical approach), reduced SNPs (reduced classical approach) and QTL-SNPs. QTL-SNPs are SNPs which were within or close to known QTL regions, including the two known QTN. Enabling a direct and fair comparison between the different approaches, a special evaluation design was applied, i.e., an invariable double 10-fold cross-validation design. The main focus in this part was to compare the observed prediction precisions of the different approaches. A second objective was to compare positions of selected important SNPs from the metabolite approach as well as from the reduced classical approach with known QTL positions using enrichment analysis as a possible way to prove their relevance for the investigated milk trait.

4.2 Material and Methods

4.2.1 Conceptual comparison between simulated and experimental data

In this section the settings for the purely conceptual comparison between experimental data set (based on Btau4.0) and a simulated data set for the conventional approach as well as the SBML approach is presented. For this comparison, the analysis framework conditions were designed to be as similar as possible for the analysis of the experimental data set and the simulated data sets. The underlying construction of the genome of the simulated data was based on the experimental data (cf. Section 3.2.1.1 on page 46). Hence, similarity was observed for the obtained LD between SNPs over the whole genome for simulated and experimental data (cf. Figure 3.5 on page 59). The main difference between the experimental data and simulated data is that the underlying numbers of QTL (i.e., the underlying genetic) are unknown in experimental data and well known in simulated data. At this point, however, we don't want to make statements about which approach for simulation of data is more appropriate in view of the experimental data, rather the objective is to compare the genetic architecture between simulated and experimental data. A more direct comparison between simulated and experimental data seems to be not possible at this stage.

For the conceptual comparison we used the experimental data set based on the Btau4.0 map (cf. Section 2.2.1 on page 31), which comprises 1,307 Holstein Friesian cows and $n_{SNP} = 43{,}079$. The following milk traits were chosen: fat content, casein content and pH value. Milk traits were standardized and the following linear model was fitted (similar to a test-day model, Ptak and Schaeffer, 1993):

$$y_{ijk}^{milk\ trait} = ah_i \times stp_j + b_1 \cdot ltp + b_2 \cdot ltp^2 + \epsilon_{ijk}, \tag{4.1}$$

with

$y_{ijk}^{milk\ trait}$ = vector of observed milk trait $[y_{ijk}^{milk\ trait} = (y_{ijk,\ 1}^{milk\ trait},...,y_{ijk,\ n}^{milk\ trait})']$,
ah_i = farm $(i = 1,...,18)$,
stp_j = test-day $(j = 1,...,39)$,
ltp = day of lactation $(ltp \in \{21,...,120\})$,
ϵ_{ijk} = residuals $(k = 1,2,...)$.

As fixed effects we considered the interaction of farm and test-day (63 levels), and the linear and quadratic regression on day of lactation in order to account for variations in the composition of milk in different stages of lactation, where b_1 and b_2 are the regression coefficients. Residuals were considered as independent and normally distributed $\epsilon_{ijk} \sim N(0,\sigma_\epsilon^2)$. The obtained residuals (ϵ_{ijk}=$y_{ijk}^{corrected}$) were used for further analyses. For comparative purposes, the narrow-sense heritability (h^2) was estimated based on the sire model (mixed model), using the R package nlme (Pinheiro et al., 2009; R Development Core Team, 2010). The sire model included the same fixed effects as modeled in Eq. 4.1 plus a random effect for sires (se_l, $l = 1,...,214$) to account for similarities among half-sibs. The use of the sire model enables the estimation of the sire variance ($\hat{\sigma}_s^2$) as well as the variance of residuals. In this context, the following relation between additive genetic variance (σ_a^2) and variance of sire exists (Falconer and Mackay, 1996, pp. 167-169):

$$\sigma_s^2 = \frac{1}{4}\sigma_a^2, \tag{4.2}$$

which means within a half-sib family all progenies share on average 25% of their genes of the sire. Thus h^2 can be estimated as follows (cf. Section 1.2 on page 5; Falconer and Mackay, 1996, pp. 167-169):

$$\hat{h}^2 = \frac{4\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_e^2}. \tag{4.3}$$

For our selected milk traits we obtained: $\hat{h}^2 = 0.234$ for fat content, $\hat{h}^2 = 0.238$ for casein content, and $\hat{h}^2 = 0.392$ for pH value.

To allow a conceptual comparison between milk traits and simulated data sets, we chose a simulated training data set (created as described in Section 3.2.2.2 on page 49) wherein 1,307 animals were randomly selected for the conventional approach as well as for the SMBL approach. The following settings were chosen for simulated data sets: $H^2 = 0.3$, $n_{SNP} = 52{,}273$ and $n_{QTL} \in \{23, 230\}$. Investigations were limited to $H^2 = 0.3$, because the chosen milk traits had similar values of $\hat{h}^2$. All analyses were realized using the fastBayesB method (as described in Section 3.2.2.3 on page 53), in which additive and dominance genetic effects are considered (cf. Eq. 3.7 on page 54). The following set of plausible values for $\gamma \in \{0.1,0.05,0.025,0.01,0.005,0.001,10^{-4},10^{-5}\}$ was tested for each phenotype (simulated and experimental). The prediction precision was obtained by using a 10-fold cross-validation (Hastie et al., 2009, pp. 241-249), for which the whole

data set was divided into 10 equally sized training sets and corresponding test sets (see Section 4.2.2). In addition, the experimental data set was divided assuring equal proportions of half-sib families. This implementation of a cross-validation approach was followed for the sake of comparability, because no separate experimental test set was available. In this context, for a simulated phenotype ($\rho_{simulated}$) the prediction precision is defined as the correlation between estimated genetic values ($\hat{g}^{simulated}$) and simulated phenotype ($y^{simulated}$), and can be expressed as:

$$\rho_{simulated} = cor(\hat{g}^{simulated}, y^{simulated}), \qquad y^{simulated} \in \{y^{conv}, y^{sbml}\}. \tag{4.4}$$

For an investigated milk trait ($\rho_{experimental}$) the prediction precision is defined as the correlation between estimated genetic values ($\hat{g}^{experimental}$) and obtained residuals ($y_{ijk}^{corrected}$), and can be expressed as:

$$\rho_{experimental} = cor(\hat{g}^{experimental}, y_{ijk}^{corrected}). \tag{4.5}$$

In addition, the square root of the estimated heritability ($\hat{h}^2$) for a trait can be used as a possible upper bound for the prediction precision which can be obtained for the investigated phenotype, since the following relation is known: $h = \rho$ (Falconer and Mackay, 1996, pp. 160-161). Also, the goodness of model fit was evaluated visually for the whole data set involving all investigated phenotypes (simulated and experimental). In the next section, the 10-fold cross-validation will be explained in more detail and also the resulting double 10-fold cross-validation design which is used in Section 4.2.3 and Section 4.2.4.

4.2.2 Cross-validation designs

For all presented analyses in this chapter, either a 10-fold cross-validation (Hastie et al., 2009) or a double 10-fold cross-validation design was applied. Which kind of cross-validation design was used is specified in the corresponding sections. In general, a cross-validation approach was necessary, because we did not have a separate experimental data set as test set available, and thus to enable investigations on the experimental data set it was divided as described in the following. Figure 4.1 illustrates a schematic representation of the applied double 10-fold cross-validation design.

10-fold cross-validation design: The whole data set was divided into 10 equal parts with equal proportions of half-sib families. To create a corresponding training set for a test set, the remaining outer test sets were merged. In detail, to create training set No.1 for test set No.1, the following test sets were combined: test set No.2 (Part2) to test set No.10 (Part10). This was realized for each test set and thus each cow appeared exactly once in each test set.
This design allows to estimate genetic effect sizes or metabolite importances for a milk

trait in the training set which are then applied to the corresponding test set to obtain the prediction precision.

Double 10-fold cross-validation: First, to obtain the outer 10-fold cross-validation, the data set was divided as described for 10-fold cross-validation design. Here the test sets are termed outer test sets and training sets are termed outer training sets. The inner 10-fold cross-validation was obtained by dividing each outer training set into 10 equal parts representing the inner test sets, and the corresponding inner training sets were assembled as explained above for the outer training sets.
The inner cross-validation of the design was used for example for parameter optimization or find important metabolites for an investigated milk trait for the proposed metabolite approach, whereby the outer cross-validation was only applied to obtain the prediction precision.

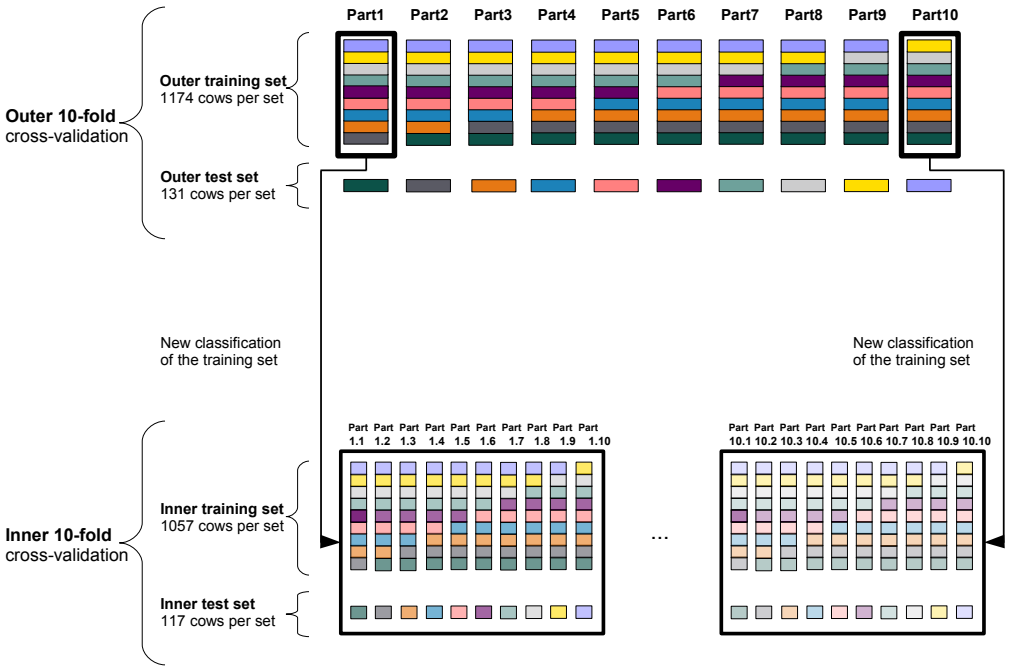


Figure 4.1: Scheme of the invariable double 10-fold cross-validation design.

4.2.3 Investigations of milk metabolites and milk traits

In this section, the different statistical analysis methods to investigate the milk metabolites and milk traits as well as the relationship between metabolites and milk traits are presented for the analysis of the experimental data set. These analyses were based on the marker map Btau4.2. Before the analyses are presented, further information for the experimental data set are provided. The data set comprises 1,305 Holstein Friesian cows, where on average, each sire had 6 daughters (range between 1 and 112 cows), each farm

had 73 cows (range between 36 and 135 cows) and each lactation day had 13 cows (range between 1 and 27). The values of day of lactation were grouped into 10-day intervals, termed lactation interval, resulting in the following intervals: lactation days 21-30 = lactation interval 1, lactation days 31-40 = lactation interval 2, and so on, up to lactation days 111-120 = lactation interval 10. On average, 131 milk samples (ranging from 55 to 183) were analyzed per lactation interval.

4.2.3.1 Statistical model for metabolites and milk traits

The following statistical mixed model was fitted to milk traits:

$$y_{ijkl}^{milk} = (ah_i \times stp_j + b_1 \cdot ltp + b_2 \cdot ltp^2) + se_k + \varepsilon_{ijkl}, \tag{4.6}$$

with

$$\begin{aligned}
y_{ijkl}^{milk} &= \text{vector of observed trait } [y_{ijkl}^{milk} = (y_{ijkl\ 1}^{milk},...,y_{ijkl\ n}^{milk})'], \\
ah_i &= \text{farm (i=1,...,18)}, \\
stp_j &= \text{test-day (j = 1,...,39)}, \\
ltp &= \text{day of lactation } (ltp \in \{21,...,120\}), \\
se_k &= \text{sire } (k = 1,...,214), \\
\varepsilon_{ijkl} &= \text{residuals } (l = 1,2,...)
\end{aligned}$$

and the following statistical mixed model was fitted to metabolites:

$$y_{ijklm}^{met} = (ah_i \times stp_j + b_1 \cdot ltp + b_2 \cdot ltp^2 + gld_m) + se_k + \tilde{\varepsilon}_{ijkml}, \tag{4.7}$$

with

$$\begin{aligned}
gld_m &= \text{GC-MS batch } (m = 1,...,47), \\
\tilde{\varepsilon}_{ijkml} &= \text{residuals } (l = 1,2,...),
\end{aligned}$$

and all other parameters are defined as in Eq. 4.6.
In both presented statistical models, the interaction ($ah \times stp$) farm and test-day (63 levels) was considered as a fixed effect (in parentheses), additionally GC-MS batch for metabolite levels. Linear and quadratic regression on day of lactation (cf. Section 4.2.1) was considered to model changes during lactation, where b_1 and b_2 were the regression coefficients. As a random effect the sire effect was considered with $se_k \sim N(0,\sigma_s^2)$ and accounted for the half-sib structure in both statistical models (note both models represent a sire model). Based on the pedigree data received from vit Verden, 192 sires could be assigned to the cows, and 22 cows had unknown sires. Residuals were assumed to be independently and normally distributed $\varepsilon_{ijkl} \sim N(0,\sigma_\varepsilon^2)$ and $\tilde{\varepsilon}_{ijkml} \sim N(0,\sigma_{\tilde{\varepsilon}}^2)$.
Depending on the analysis the standardized residuals ($y_{corrected}^{milk}$) of the milk traits were

used, and determined as follows:

$$y^{milk}_{corrected} = y^{milk}_{ijl} - \delta, \tag{4.8}$$

and the standardized residuals ($y^{met}_{corrected}$) for metabolites were determined as follows:

$$y^{met}_{corrected} = y^{met}_{ijml} - \tilde{\delta}, \tag{4.9}$$

where δ represents the obtained fitted fixed effects for milk traits from Eq. 4.6 and $\tilde{\delta}$ for milk metabolites from Eq. 4.7. From this follows that the sire effect was only used to improve the estimation of the fixed effects for correction.

4.2.3.2 The impact of influencing factors on investigated traits

The impact of an influencing factor on metabolite profiles or on milk traits was studied with the following statistical tests: For the fixed effects, an F-test (ANOVA) was applied in sequence to all traits. For the random effect, a one-sided likelihood ratio test (LRT) was applied. The testing problem was H_0: $\sigma^2_s = 0$ versus the alternative hypothesis H_A: $\sigma^2_s > 0$. The distribution of the LRT statistic under the null hypothesis approximately followed a mixture of χ^2-distributions according to Self and Liang (1987). To investigate a specific influencing factor, metabolite measurements were corrected over all animals for all influencing factors except for the one of interest following Eq. 4.6 for milk traits and Eq. 4.7 for metabolites. The observed P-values were corrected, because of multiple testing, using the false discovery rate (FDR) controlling method by Benjamini and Hochberg (1995), and fixing the estimated FDR at 5%. We applied the FDR correction as implemented in the R package multtest (Pollard et al., 2010). This correction method was applied to all tests (e.g., for each influencing factor).
To investigate if metabolite intensities significantly differed between levels of an influencing factor of interest, standardized residuals (cf. Eq 4.9) were used. To compare means pairwise for levels of categorical factors for a metabolite, we applied the Tukey test (Kramer, 1956), using the R package DTK (Lau, 2011). Also, for day of lactation it was possible to visually prove if an increase or decrease trend existed for metabolites over lactation days, based on the ANOVA test. For this, the estimated regression coefficients $\hat{b_1}$ and $\hat{b_2}$ from the full statistical model (cf. Eq. 4.7) were applied to estimate the metabolite, $\hat{y}^*_{met}$, as follows:

$$\hat{y}^*_{met} = \hat{b_1} \cdot ltp + \hat{b_2} \cdot ltp^2. \tag{4.10}$$

All analyses were implemented using R (R Development Core Team, 2010).

4.2.3.3 Multivariate analyses investigating specific influencing factors on milk metabolites

To investigate a specific influencing factor, metabolite measurements for all animals were corrected for all influencing factors except for the one of interest according to the mixed model in Eq. 4.7, and the standardized residuals ($y^{met}_{corrected}$) were used. The following analyses were applied on the influencing factor farm and day of lactation, since both have an impact on the metabolic state of the cow.

Clustering of metabolite profiles regarding influencing factors: For this purpose, the mean over all samples on a specific level of the influencing factor under investigation was taken for each metabolite. The Euclidean distance between vectors of mean metabolite measurements was used to determine similarities between metabolite profiles for the levels of the investigated influencing factor. Hierarchical clustering using the method of average linkage was applied. Two validation criteria were used to evaluate the number of clusters which had to be determined in advance. The silhouette width criterion (Rousseeuw, 1987; Vendramin et al., 2009) provided information about compactness and separation of clusters. The stability of clusters (Hennig, 2007) was calculated using the function clusterboot (R package FPC; Hennig, 2010). This function assesses the clusterwise stability of clustering resampled data; the number of bootstrap rounds was 1,000. Therein, the Jaccard coefficient (Jaccard, 1901; Vendramin et al., 2009) was used as a similarity measure.

Classifying levels of influencing factors: The linear discriminant analysis (LDA) (Fisher, 1936; Hastie et al., 2009, pp. 106-112), using the R package MASS (Venables and Ripley, 2002), was used to investigate multivariate relations of standardized metabolite profiles to the influencing factor of interest. On the one hand, the estimated discriminant function was used to determine important metabolites for specific factors as follows. A 10-fold cross-validation was implemented (cf. Section 4.2.2 on page 77). In each cross-validation run, the coefficients of the first linear discriminant function, which explains most of the between-group variance, were recorded for all metabolites. The coefficients were used as a measure of association between each milk metabolite and the investigated influencing factor. We defined a metabolite to be important if its coefficient was larger than the 90% quantile of the absolute coefficients for all metabolites in each cross-validation run.
On the other hand, the estimated discriminant function was also used to classify new data taking the 10-fold cross-validation approach, to quantify the strength of association and its significance. The precision of prediction was determined as the proportion of correctly classified samples reporting the mean, based on ten cross-validation runs. To quantify significance of the observed prediction ability, we applied a resampling approach to randomly destruct the possible association between factor levels and metabolite profiles. The number of resampling rounds was 1,000, and in each resampling round again a

10-fold cross-validation was applied (cf. Section 4.2.2). The resampling P-value of the observed prediction precision was determined as usual for a permutation test (Good, 2005) as the relative proportion of resampling rounds with a prediction precision as large or larger than the observed precision of prediction for the original data.

4.2.3.4 Analyses of milk traits related to metabolites

In total 14 milk traits were investigated, where 11 milk traits were measured in the MPT, and three additional milk traits characterizing the status of EB: ratio fat:protein, ratio fat:lactose and energy content of milk. Energy content of milk (MJ/kg) was determined as follows (Kirchgessner, 1992, p. 284):

$$Energy = 0.39 \cdot fat\ \% + 0.24 \cdot protein\ \% + 0.17 \cdot lactose\ \%. \tag{4.11}$$

SCC was transformed to somatic cell score (SCS) following Ali and Shook (1980):

$$SCS = log_2 \frac{SCC}{1000} + 3. \tag{4.12}$$

All investigated milk traits as well as the corresponding estimated heritabilities using the full statistical mixed model (cf. Eq. 4.6) are listed in Table 4.1. Milk metabolites and milk traits were corrected for the influencing factors as modeled in Eq. 4.6 and Eq. 4.7 on page 79. The observed residuals were standardized and used in the subsequent analyses presented here.

The R function cor.test was applied, using Pearson's correlation coefficient, to test correlations between paired samples (ρ_{t1t2}) of milk metabolites and milk traits, between

Table 4.1: Estimated heritabilities for investigated milk traits.

Milk trait	$\hat{h}^2$
Acetone (%)	0.207
Casein (%)	0.240
Fat (%)	0.233
Lactose (%)	0.082
pH value	0.387
Protein (%)	0.240
SCS	0.087
SFA	0.216
Quantity of milk (kg)	0.151
UFS	0.121
Urea (%)	0.163
Fat:protein	0.168
Fat:lactose	0.238
Energy (MJ/kg)	0.252

milk traits and between milk metabolites.
The regression methods RF (Breiman, 2001) and partial least squares (PLS; Wold, 1975; Hastie et al., 2009, pp. 80-82) were used to predict a milk trait from all corrected metabolite profiles. The procedures were used as implemented in the R packages randomForest (Liaw and Wiener, 2002) and mixOmics (Dejean et al., 2011) for PLS. Further, a 10-fold cross-validation was implemented to determine the precision of prediction (ρ_{milk}), which is defined as the correlation between predicted and observed values of a milk trait, and can be expressed as:

$$\rho_{milk} = cor(y^{milk}_{corrected}, \hat{y}^{milk}_{corrected}). \tag{4.13}$$

Additionally, for PLS it is necessary to determine the number of latent components to achieve a minimal prediction error, which was measured as the mean squared error of prediction. To determine an optimal number of latent components for prediction in each (outer) cross-validation run, an inner 10-fold cross-validation was implemented (cf. Section 4.2.2). The vip function of the R package mixOmics was used to extract the metabolite importance for PLS. In RF, after finishing the (outer) 10-fold cross-validation runs, a resulting mean decrease in accuracy was used as a measure of importance of metabolites. To determine the metabolites important for predicting a milk trait for each prediction method, we used the 90% quantile of the importance measurements of all metabolites in each cross-validation run, and defined a metabolite to be important if its importance measurement was larger than the 90% quantile in each of the ten cross-validation runs.

4.2.4 The metabolite approach for experimental data

In this section the realization of the metabolite approach, classical approach, reduced classical approach and QTL approach are presented, resulting in different SNP subsets used for the genetic value prediction. The different approaches are compared for the following milk traits: fat content, protein content and pH value.
The experimental data set based on the marker map Btau4.2 was used for this comparison. We used the following information from 1,305 Holstein Friesian cows: 40,317 SNPs, 190 milk metabolites and 11 milk traits (cf. Section 2.2.4 on page 36). In the next section the filtering steps to obtain the SNP subset for the QTL approach are presented.

4.2.4.1 Known QTL regions for fat and protein

For the implementation of the QTL approach, the cattle QTL database (cattleQTLdb; http://www.animalgenome.org/cgi-bin/QTLdb/BT/index; Hu et al. (2007)) was searched to determine known QTL regions of the bovine genome based on the given cattleQTLdb markers for fat and protein. Entries of the cattleQTLdb were filtered for: trait milk fat percentage and milk protein percentage, analysis type equal to QTL,

breed equal to Holstein, and chromosome number, flanking markers (of the confidence interval of the QTL) or peak markers had to be specified. The location of selected cattleQTLdb markers is given in the genetic unit cM. Then, these markers were assigned to Btau4.2 (as used for the experimental data), using the corresponding marker information from the National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nih.gov//genomes/MapView/Bos_taurus/sequence/BUILD.5.2/initial_release/, Btau_4.2-Primary Assembly), to obtain marker positions in the physical unit bp. In total, 34 QTL regions were associated with fat, and 50 QTL regions with protein. The QTL marker positions used for both milk traits are listed in the Appendix B.2 on page 151. Additionally, the known QTN DGAT1 (Grisart et al., 2004) was considered a QTL for fat and protein. The position of another known QTN for protein, ABCG2 (Cohen-Zinder et al., 2005), was already covered by a QTL. Based on the filtered QTL marker positions (bp) it was possible to select SNPs close to a QTL peak marker or between two flanking markers of a QTL region:

(a) QTL region: all SNPs between left (end bp) and right (start bp) flanking marker of the QTL interval.

(b) QTL peak: left and right SNP next to the peak position.

(c) DGAT1: a SNP directly in the DGAT1 region (DGAT1: chromosome 14, position 411,147-446,810 bp, NCBI accessed Feb. 2011). The SNP is located on 443,937 bp and termed DGAT1-SNP.

The joint set of SNPs (a)-(c) were termed QTL-SNPs in the following analyses.

4.2.4.2 Analysis design to investigate the different approaches

The following three-step analysis design was performed to investigate associations between three levels of data: SNP-genotypes, milk metabolites and milk traits. Figure 4.2 illustrates the different approaches: metabolite, classical, reduced classical and QTL approach. Realization based on a double 10-fold cross-validation.

Step 1: The standardized residuals of milk traits (cf. Eq. 4.8 on page 80) and metabolites (cf. Eq. 4.9 on page 80) were used for the regression of milk traits on metabolite profiles with RF and PLS as described in Section 4.2.3.4. Here, a metabolite was defined as important for a specific milk trait if its measure of importance was larger than the 90% quantile of all metabolite importances in each inner cross-validation run and for each regression method. In this step, the prediction precision (ρ_{milk}) was defined as correlation between predicted and observed milk trait values (cf. Eq. 4.13 on page 83). Analyses were implemented in R (R Development Core Team, 2010). This step was only realized for the metabolite approach.

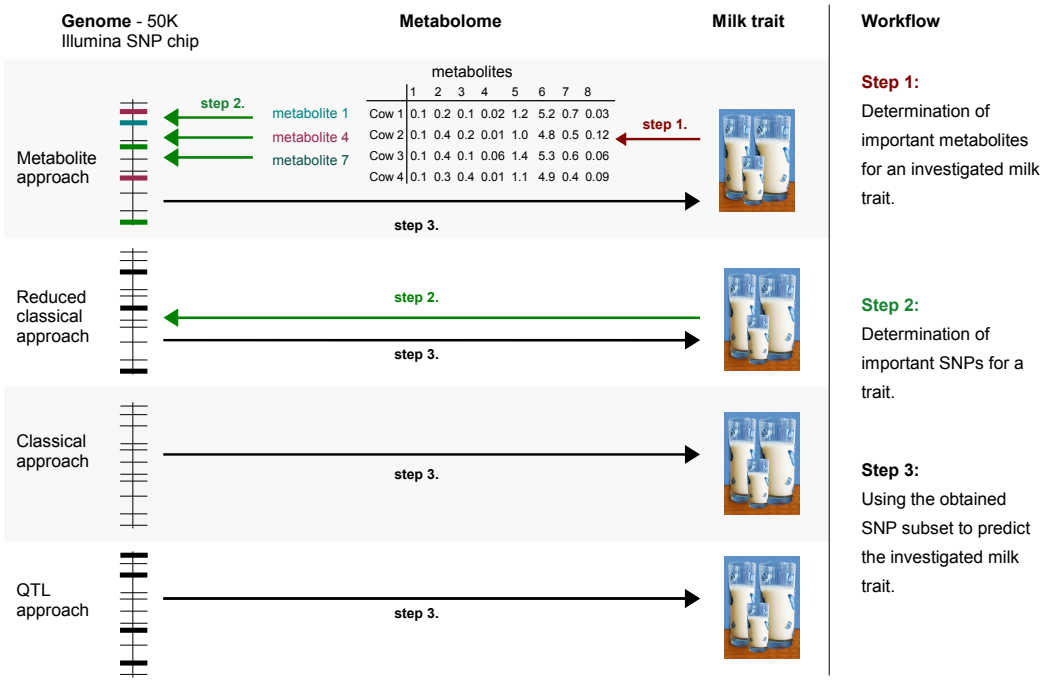


Figure 4.2: Analysis design: schematic representation of the workflow. In the classical approach all SNPs were used to predict the genetic values, whereas only the thick marked SNPs were used for the genetic value prediction in all other approaches.

Step 2: The impact of each SNP on either the important metabolites or the milk traits was estimated using the outer training set. An SVS method similar to Ishwaran and Rao (2005) was applied (Wittenburg and Reinsch, 2011), including the estimation of the fixed effects and additive genetic effects covered by SNPs. Thus milk metabolites and milk traits were only standardized before they were used.
The used model for milk traits:

$$y^{milk}_{corrected} = Ff + X\tilde{a} + \tilde{\varepsilon}, \tag{4.14}$$

with

$$\begin{aligned}
y^{milk}_{corrected} &= \text{vector of investigated milk trait } [y^{milk}_{corrected} = (y^{milk}_{corrected\ 1},\ldots,y^{milk}_{corrected\ n})'],\\
F &= (n \text{ x } f)\text{-design matrix for fixed effects},\\
f &= \text{fixed effects as modeled in Eq. 4.6 presented in parentheses } f = (f_1,\ldots,f_{66})'),\\
X_{\tilde{a}} &= (n \text{ x } n_{SNP})\text{-design matrix for additive genetic effects } \tilde{a} = (\tilde{a}_1,\ldots,\tilde{a}'_{n_{SNP}}) \ ,\\
\tilde{\varepsilon} &= \text{residuals},
\end{aligned}$$

and for milk metabolites:

$$y^{met}_{corrected} = F^* f^* + Xa^* + \varepsilon^*, \tag{4.15}$$

with

$y^{met}_{corrected}$ = vector of investigated milk metabolite [$y^{met}_{corrected} = (y^{met}_{corrected\ 1},...,y^{met}_{corrected\ n})'$],
F^* = (n x f^*)-design matrix for fixed effects,
f^* = fixed effects as modeled in Eq. 4.7 presented in parentheses $f^* = (f^*_1,...,f^*_{113})'$,
Xa^* = (n x n_{SNP})-design matrix for additive genetic effects $a^* = (a^*_1,...,a^*_{n_{SNP}})'$,
ε^* = residuals.

The residuals were assumed to be independently and normally distributed $\tilde{\varepsilon}_i \sim N(0,\sigma^2_{\tilde{\varepsilon}})$ or $\varepsilon^*_i \sim N(0,\sigma^2_{\tilde{\varepsilon}^*})$. Entries of the design matrices are random variables and depends on the observed SNP-genotypes. SNP-genotypes in the design matrices F and F^* are coded as presented in Section 3.4 on page 50, whereby the homozygous with the most frequent allele is coded as one. It is assumed that genetic effects are independently distributed at different loci. Here LE is not explicitly required. The following assumptions of prior distributions were used (following Wittenburg and Reinsch, 2011) and the additive genetic effects were standardized:

$$\tilde{y}|g_s,\sigma^2_{res} \sim N(\tilde{X}g_s, I\sigma_{res}), \quad \tilde{y} \in \{y^{milk}_{corrected},y^{met}_{corrected}\},\ g_s \in \{\tilde{a},a^*\},\ res\ \in \{\tilde{\varepsilon},\varepsilon^*\}$$
$$\tilde{X} \in \{F,F^*\},\ I = \text{identity matrix},$$
$$\sigma^{-2}_{res}|\beta_1,\beta_2 \sim \Gamma(\beta_1,\beta_2),$$
$$g_{s,j}|\sigma^2_{s,j} \sim N(0,\sigma^2_{s,j}), \quad j \in \{1,...,n_{SNP}\},$$

where $\sigma^2_{s,j}$ is drawn from a mixture of inverse Γ-distributions. The mixture depends on the complexity parameter ω, which gives the proportion of effects different to zero. The following settings of parameters were used (following the notation of Wittenburg and Reinsch (2011)): $\beta_1 = \beta_2 = 0.00001$, $\alpha_1 = 5$, $\alpha_2 = 0.01$ and $\upsilon_o = 0.001$ for SVS (for more information see Wittenburg and Reinsch, 2011; Ishwaran and Rao, 2005). SVS was run using Gibbs sampling with the following settings: 100,000 iterations were used, the first 40,000 of which were disregarded as burn-in phase. Example trace plots for selected SNPs and for each milk trait are presented in Appendix B.3 on page 156. Three chains were produced for each trait. To enable the determination of significant SNPs, a conditional test was used similar to that used in Wittenburg and Reinsch (2011). The testing problem was H_0: $g_{s,j} = 0$ versus the alternative hypothesis H_A: $g_{s,j} \neq 0$. Therefore, an empirical selection method was applied. In each chain was counted (after the burn phase) how often a marker has non-zero effect. The mean value of estimated marker effects was used in combination with the mean estimated complexity parameter ω to select important SNPs. The additive genetic effect sizes as mean of estimates were used for the prediction of the specific investigated trait (in Step 3).

After this step was completed, the important SNPs were rated related to the known

QTL using the QTL-SNPs. For this, we used the over-representation analysis which is a type of enrichment analysis (Ackermann and Strimmer, 2009). The aim of this analysis was to determine if a list of genes, representing the gene set, is over-represented (more genes than expected by chance) with regard to another gene list, representing the target set. A specific reference set is applied to quantify how likely the over-representation is, which is calculated following the hypergeometric distribution (Drăghici et al., 2003). In our case, the entirety of SNPs represent the reference set. The target set corresponds to the QTL-SNPs, and we investigated its enrichment with regard to the SNP subsets detected in the metabolite approach and the reduced classical approach. The analysis was performed in R (R Development Core Team, 2010), using the function phyper to calculate the P-values based on the hypergeometric distribution. The significance level α was set to 0.05.

Step 3: Different SNP subsets were used to estimate genetic effects on milk traits using SVS (same settings as in Step 2):

(a) metabolite SNPs,

(b) reduced SNPs,

(c) all SNPs and

(d) QTL-SNPs.

Here, additive genetic effects as mean of estimates (in the outer training set) were used to predict the genetic values in the outer test set. In this step, the prediction precision (ρ_{svs}^{milk}) was defined as the correlation between the estimated genetic values ($\hat{g}_{svs}^{milk}$) and the observed characteristics of a milk trait (y^{milk}), and can be expressed as:

$$\rho_{svs}^{milk} = cor(\hat{g}_{svs}^{milk}, y^{milk}). \tag{4.16}$$

Finally, the prediction precisions (ρ_{svs}^{milk}) of the four different approaches were rated using Wilcoxon signed-rank test for paired samples. The analysis was performed in R (R Development Core Team, 2010), using the function wilcox.test (α was set to 0.05). The rating allows us to determine if the obtained prediction precisions differed significantly among the various investigated SNP subsets for SVS.
This was possible due to the invariability of the used double 10-fold cross-validation scheme (cf. Section 4.2.2) and also the use of the same seeds for the random number generator in the analyses, which ensured the comparability of the different approaches.

To evaluate the significance of the prediction results for the reduced classical approach and the metabolite approach, it was tested if SNP subsets determined for a milk trait were superior to random subsets. To quantify the significance of the observed prediction ability for the original design, we applied a resampling approach for which SNP subsets

were chosen randomly for each investigated milk trait. For each of the 10 outer cross-validation runs 100 SNP subsets were drawn at random corresponding to the observed average quantity of SNPs in the respective approach and step 3 was processed. Thus, the evaluation was based on 1,000 resampling rounds, resulting in an empirical distribution of prediction precisions (ρ_R). The resampling P-value of the prediction precision was determined in the same way as described in Section 4.2.3.3 (on page 81) as the relative proportion of resampling rounds with a prediction precision ρ_R as large or larger than the original prediction precision (ρ_{svs}^{milk}; α was set to 0.05).

4.3 Results of investigations of the experimental data set

4.3.1 Analysis of experimental and simulated data

Two different approaches to simulate data were conceptually compared to experimental data by comparing the results of fastBayesB analyses. We observed that the optimal γ-value disagreed for the different investigated experimental data sets, e.g., pH value $\gamma = 0.001$, fat content $\gamma = 10^{-4}$ and casein content $\gamma = 10^{-5}$. The estimated variance components and prediction precisions for the optimal γ-value can be found in Table 4.2. In general, H^2 was underestimated by fastBayesB compared to estimated h^2 obtained with the sire model (cf. with page 75) for all investigated milk traits; for example, $\hat{h}^2 = 0.234$ with the sire model and $\hat{H}^2 = 0.105$ with fastBayesB for fat content. Furthermore, the observed mean prediction precisions can be found in this Table (cf. Eq. 4.4 on page 77) which was scaled to 100% to ease conceptual comparison between simulated and experimental data sets. Therefore, the prediction precisions were divided by the square root of the estimated or simulated heritability (cf. Section 4.2.1 on page 75). In Figure 4.3 the estimated genetic effect sizes are presented for the different milk traits observed using the whole data set. In this figure, it is shown that casein content revealed

Table 4.2: Estimated variance components and prediction precisions for simulated data sets and milk traits with fastBayesB. For comparison, an example training data set was selected including 1,307 animals (settings: $n_{SNP} = 52{,}273$ and $H^2 = 0.3$) for each kind of simulated data set. The experimental data set included 1,307 animals, and three milk traits were studied. 10-fold cross-validation was applied to determine prediction precision. The average variance components and in brackets the corresponding standard deviation are given as well as the prediction precisions are presented for the optimal γ-value.

Data	Approach	$\hat{\sigma}_g^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_d^2$	$\hat{\sigma}_e^2$	$\hat{H}^2$	ρ	scaled ρ
n_{QTL}=23	conventional	0.714 (0.07)	0.627 (0.06)	0.087 (0.02)	2.561 (0.03)	0.218	0.454 (0.07)	82.36%
	SBML	0.756 (0.09)	0.756 (0.09)	0.001 (0.00)	2.481 (0.06)	0.233	0.475 (0.09)	86.36%
n_{QTL}=230	conventional	0.734 (0.08)	0.640 (0.05)	0.094 (0.05)	1.839 (0.07)	0.285	0.380 (0.08)	69.09%
	SBML	0.556 (0.09)	0.554 (0.09)	0.002 (0.00)	2.676 (0.05)	0.172	0.263 (0.08)	47.82%
Experi-	Fat (%)	0.085 (0.01)	0.081 (0.01)	0.004 (0.01)	0.722 (0.02)	0.105	0.292 (0.07)	60.83%
mental	Casein (%)	0.023 (0.00)	0.023 (0.00)	0.000 (0.00)	0.674 (0.01)	0.034	0.186 (0.07)	37.96%
	pH value	0.191 (0.03)	0.143 (0.02)	0.048 (0.02)	0.345 (0.02)	0.356	0.255 (0.10)	41.12%

only one intermediate additive effect. In comparison, besides one major additive genetic effect, fat content showed three intermediate additive and one dominance genetic effects. Analyses for pH value revealed equally large genetic effects for additive and dominance effects. For simulated data sets the observed main genetic effect sizes were close to those observed using the corresponding whole training sets (see Figure 3.6 on page 60). The visual inspection of the observed fitted values and residuals using the whole data set revealed that the model explains the simulated as well as the experimental data well.

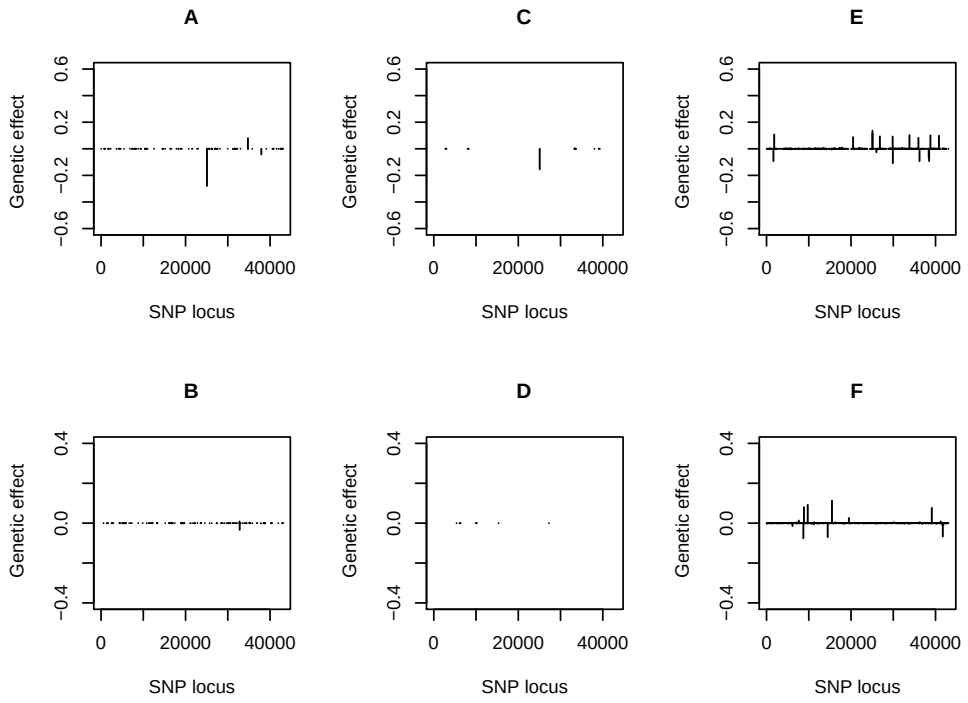


Figure 4.3: Estimated main genetic effects for different milk traits. Estimated additive (A) and dominance genetic effects (B) for fat content; additive (C) and dominance genetic effects (D) for casein content; additive (E) and dominance genetic effects (F) for pH value. The figures based on the whole data set and analyzed with fastBayesB for the optimal γ-value.

4.3.2 Analysis of milk metabolites and milk traits

In this section all analyses of the investigations of milk metabolites and milk traits are presented. Here, the focus was on the investigation of the influence of known influencing factors on milk metabolites and milk traits, as well as to study in more detail the relationships within and between milk metabolites and milk traits.

4.3.2.1 Univariate analyses of the impact of influencing factors on traits

Numbers and percentages of milk metabolites and milk traits, for which the analyzed influencing factors were significant based on ANOVA and LRT (as described in Sec-

tion 4.2.3.2 on page 80), are listed in Table 4.3. The factors showed a higher percentage of significance for milk traits than for metabolites, e.g., day of lactation had a significant influence on 45.79% of the metabolites and on 85.71% of the milk traits. More than 85% of all traditional milk traits were significantly influenced by all influencing factors. A

Table 4.3: The number of metabolites and milk traits, on which influencing factors impact significantly. Relative proportions are in parentheses.

Influencing factor	Milk metabolites		Milk traits	
Sire	34	(17.80)	13	(92.86)
Day of lactation	87	(45.79)	12	(85.71)
Farm	145	(76.32)	13	(92.86)
Test-day	105	(55.62)	12	(85.71)
Farm × test-day	159	(83.68)	14	(100.00)
GC-MS batch	190	(100.00)	-	-

detailed list of the observed corrected P-values is available online and can be found on the website of the Journal of Dairy Science (Melzer et al., 2013a). In addition, it was visually inspected if an increase or decrease trend existed for metabolites over lactation days (cf. Eq. 4.10 on page 80). We could observe that some metabolites showed an increased or decreased trend over lactation days. In Figure 4.4 A-B an example metabolite for an increased as well as a decreased trend is presented.

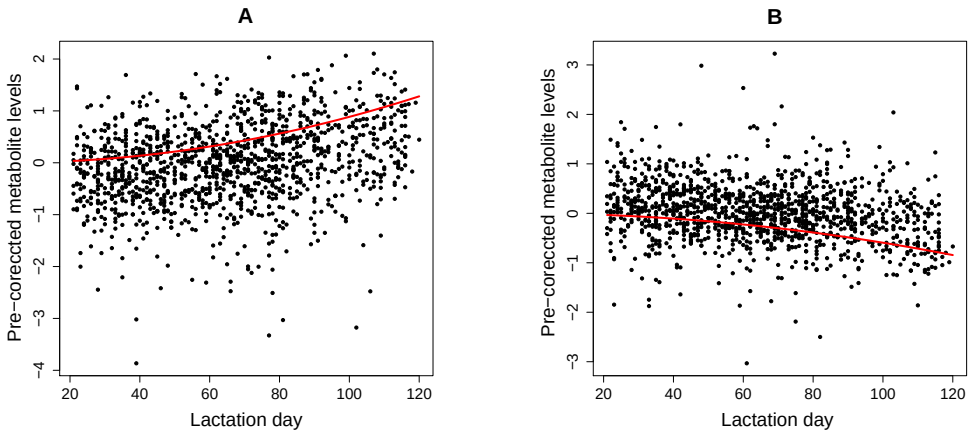


Figure 4.4: An example for an increased trend over lactation days represents Glucose-, 2-amino-2-desoxy (A) and for a decreased trend Kynureine (B).

4.3.2.2 Multivariate analyses of the impact of influencing factors on milk traits

Clustering of average metabolite profiles: In Figure 4.5, the dendrogram for the influencing factor lactation interval is presented. The lactation intervals can be split into three clusters: cluster 1 comprises lactation intervals $1-3$ (days 21-50), cluster 2 intervals $4-8$ (days 51-100) and cluster 3 intervals $9-10$ (days 101-120). The number

of clusters was evaluated using the silhouette width criterion, for which the maximum average silhouette width (asw) was determined at 0.30. For the analysis involving three clusters, the cluster stability criterion was applied, and we observed a Jaccard coefficient higher than 0.68 for all three clusters, indicating intermediate stability. The dendrogram presented reproduced the lactation time line.

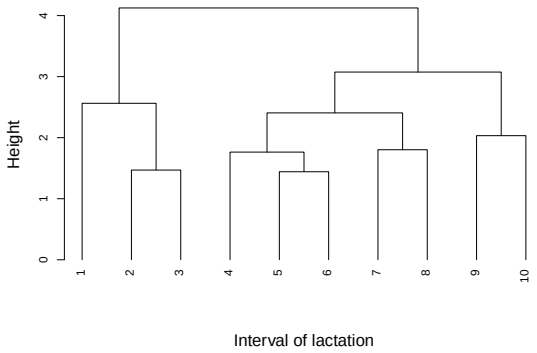


Figure 4.5: The dendrogram resulting from hierarchical clustering of average metabolite profiles for the influencing factor lactation interval. The metabolite profiles were pre-corrected using the linear model (Eq. 4.7 on page 79) except for the linear and quadratic regression on lactation day.

For the influencing factor farm the dendrogram is shown in Figure 4.6 A. The chosen number of clusters was three, which was proven by the silhouette width criterion ($asw = 0.37$) and the Jaccard coefficient, which was higher than 0.85 for all three clusters indicating high cluster stability.

Pairwise statistical tests: We tested how many metabolites showed a significant difference between levels of an investigated influencing factor. For the influencing factor lactation interval, all 45 pairwise comparisons were analyzed and the largest number of differences was found between lactation intervals 1 and 7 or 8. Metabolites with the highest numbers of significant differences for all pairwise comparisons of lactation intervals are listed in Table 4.4. For the influencing factor farm, 153 pairwise comparisons were investigated, of which 66 showed significant differences between metabolites. The number of significant differences for each farm compared to all other farms is graphically presented in Figure 4.6 B. The largest number of significant differences was found between farms 10 and 15 (16 metabolites in total). Whereas no significant metabolite was detected, e.g., between farms 6 and 12. The observed number and corresponding relative percentage of significant differences are reported for the metabolites with the highest counts in Table 4.4.

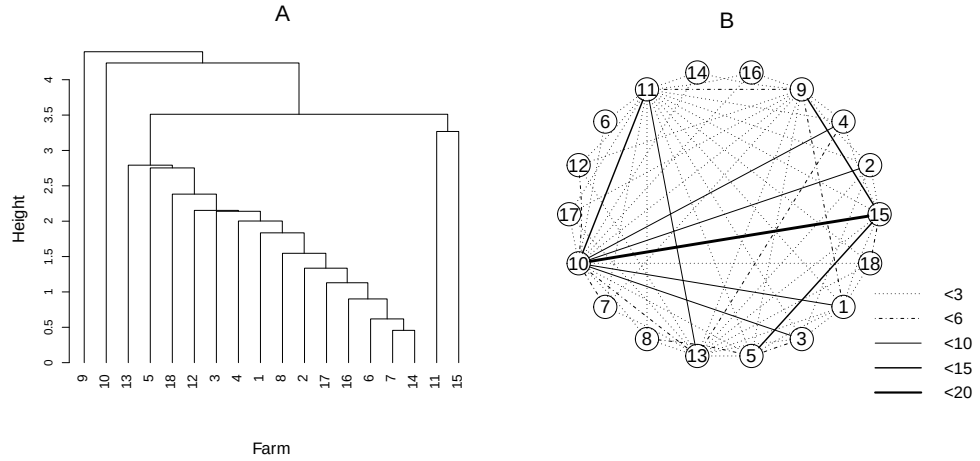


Figure 4.6: (A) The dendrogram resulting from hierarchical clustering of average metabolite profiles for the influencing factor farm. (B) Numbers of metabolites with significant differences between farms. Nodes are farms; thickness of connecting lines depends on number of metabolites with significant difference. For both graphics the metabolite profiles were pre-corrected using the full linear model (Eq. 4.7 on page 79) except for farm.

Classifying levels of specific influencing factors: We used LDA to determine important metabolites for the prediction of an influencing factor. Important metabolites were derived from the first linear discriminant function using obtained coefficients. For lactation interval, we observed 10 important metabolites, and for farm we found eight important metabolites. The most important metabolites are listed in Table 4.5 for both influencing factors. Moreover, we applied LDA to predict the influencing factor of interest from metabolite profiles. The observed precision of prediction for the original data was significant (resampling P-value ≤ 0.001) for both investigated influencing factors.

4.3.2.3 Results of investigations on relation between traits

Testing correlation: Investigating correlations between milk metabolites, we observed that in total 80% of all pairwise correlations were significant after FDR correction at a significance level of 5%. The highest correlations were found between: ethanolaminephosphate and orotic acid, and 2-methyl-fumaric acid and itaconic acid ($\rho_{t1t2} > 0.9$ for these pairs), respectively. Most metabolites were positively correlated. A detailed correlation matrix is available online and can be found on the website of the Journal of Dairy Science (Melzer et al., 2013a).

The significant correlations between metabolites and milk traits are presented in Figure 4.7. In total, 75 metabolites showed no significant correlations to any milk trait. We observed that for casein content and protein content, metabolites mostly coincided regarding significance and degree of correlation. This finding is further illustrated in

Table 4.4: The number of metabolites with the most significant differences between levels of an investigated factor of interest. The total number of possible pairwise comparisons per metabolite for each influencing factor is given in brackets. Relative proportions are in parentheses.

Influencing Factor	Milk metabolites	Number of observation	
Lactation interval [45]	Glucopyranoside, 1-O-methyl-, alpha-	26	(57.78)
	Glucosamine, N-acetyl	26	(57.78)
	Ribulose-5-phosphate	26	(57.78)
	Gluconic acid-6-phosphate	25	(55.56)
	Fructose-6-phosphate	24	(53.33)
	Phosphoenolpyruvic acid	21	(46.67)
	Sarcosine	20	(44.44)
	Galactosamine, N-acetyl-	18	(40.00)
	Arabitol	17	(37.78)
	Gluconic acid	17	(37.78)
Farm [153]	Benzoic acid	35	(22.88)
	Kynurenine	17	(11.11)
	1,3-Dihydroxyaceton	11	(7.19)
	Butanoic acid, 2-amino-	10	(6.54)
	Pyridoxal	10	(6.54)
	Thiazole, 4-methyl-5-hydroxyethyl-	10	(6.54)
	Arabitol	8	(5.23)
	Phenylalanine	8	(5.23)
	Pyruvic acid	8	(5.23)

Figure 4.7, where parts 2 and 5 of the correlation structure also show that these two traits mainly correlate with the same metabolites. Whereas Figure 4.7 part 3 shows that less congruence exists between both milk traits. Also, we observed that metabolites which were clearly positively correlated to SCS, were clearly negatively correlated to lactose content (Figure 4.7 part 3).

Investigating the correlation structure between milk traits, we found that in 81% of all pairwise correlations were significant but with very different levels of correlation, see Figure 4.8. Casein content and protein content, energy and fat content, energy and fat:lactose, fat content and fat:lactose were highly correlated ($\rho_{t1t2} > 0.9$). A positive correlation was observed, e.g., between fat content and protein content, and a negative correlation was observed, e.g., between fat content and lactose content as well as between urea content and casein content or protein content, respectively. The corresponding Pearson correlation matrix for investigations between metabolites and milk traits, and among milk traits can be found in Appendix B.1 (see page 143).

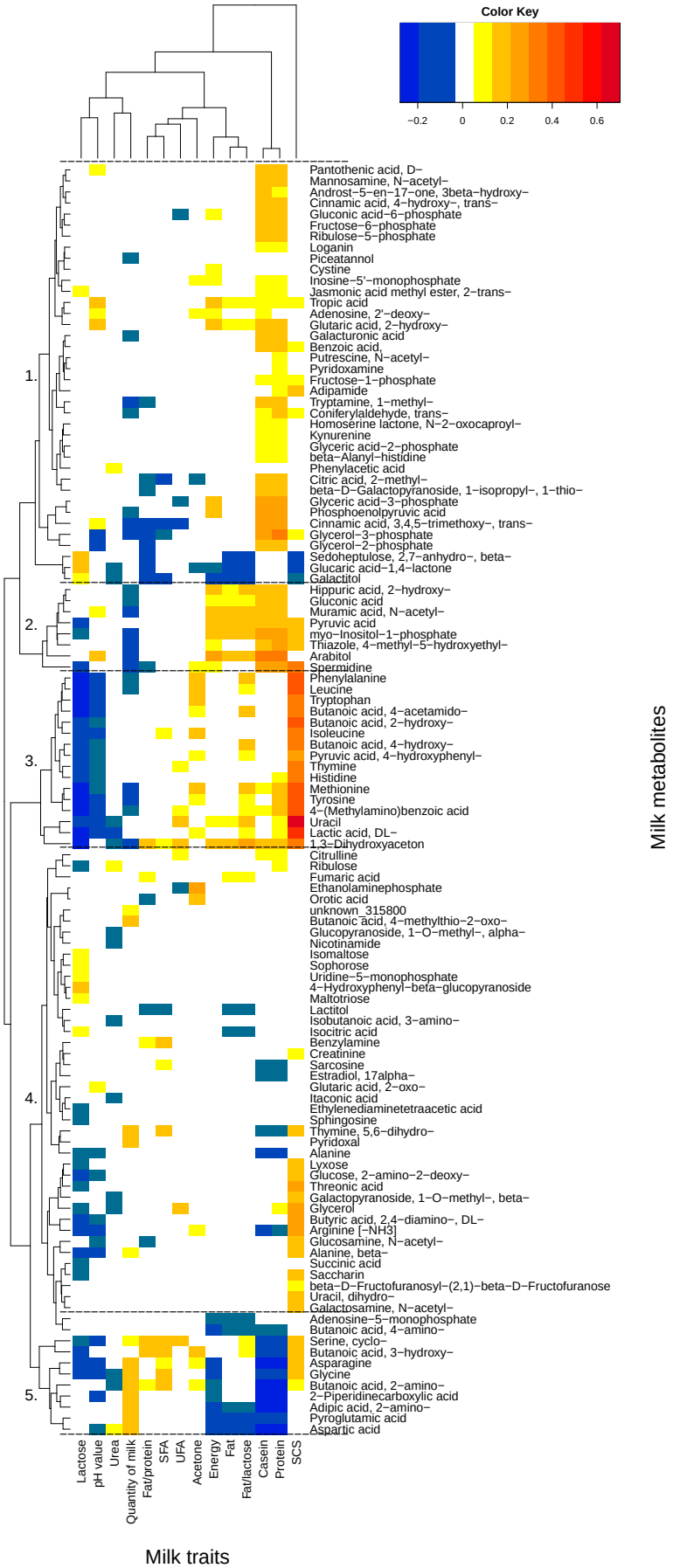


Figure 4.7: Correlations between milk metabolites and milk traits; white squares represent non significant correlations. Seventy-five metabolites had no significant correlation to any investigated milk trait and were excluded. Milk traits and metabolites were clustered using hierarchical clustering with Euclidean distance and complete linkage.

Table 4.5: Important metabolites detected for the factors lactation interval and farm, using a 10-fold cross-validation in a linear discriminant analysis. The coefficient of the first linear discriminant function was used as a measure of association between metabolites and investigated influencing factor. A metabolite was declared important if the corresponding coefficient exceeded the 90% quantile in each cross-validation run. Metabolites typed in bold were also detected in the univariate analysis (pairwise statistical test).

Lactation interval	Farm
Adipic acid, 2-amino-	**2-Piperidinecarboxylic acid**
Arabitol	Adipic acid, 2-amino-
Arginine [-NH3]	Aspartic acid
Asparagine	**Butyric acid, 2,4-diamino-, DL-**
Gluconic acid	**Galactosamine, N-acetyl-**
Glycerol	Glycerol-2-phosphate
Orotic acid	**Ribulose-5-phosphate**
Phenylalanine	**Unknown_315800**
Pyridoxal	
Unknown_315800	

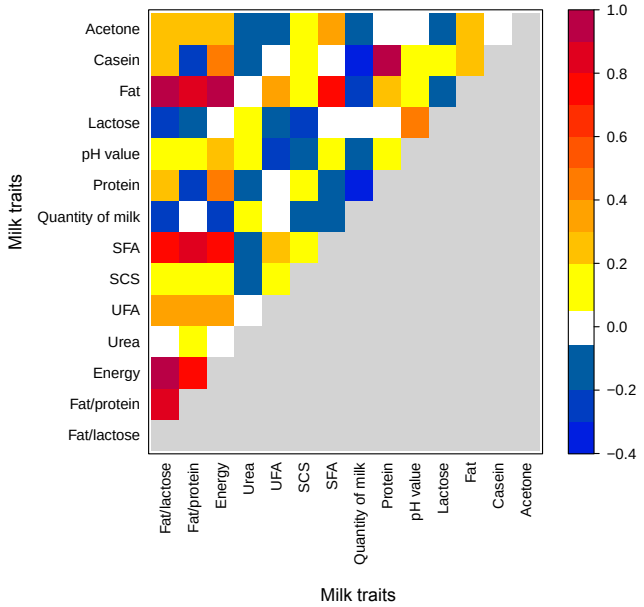


Figure 4.8: Correlations between milk traits; white squares represents non-significant correlation.

Important metabolites with multivariate relations to milk traits: The observed precisions (ρ_{milk}; Eq. 4.13) for predicting the investigated milk traits from metabolites using RF and PLS are presented in Figure 4.9. Both methods resulted in very similar precisions for the prediction of most milk traits. The highest mean value of prediction precision was observed for SCS ($\rho_{milk} = 78\%$, RF). Casein content and protein content had similar mean prediction precisions above 60%. For all other milk traits, mean prediction precision was observed between 17% and 41%. SFA had the lowest prediction precision.

For predicting SCS using RF or PLS, a very large importance was observed for one metabolite (uracil), a few metabolites showed small importance and most had nearly zero importance for this milk trait. This behavior was stable over the cross-validation runs. For fat content, more metabolites showed larger importance for both prediction methods. The results of metabolite importance measurements varied strongly over the cross-validation runs. This finding was also observed for the remaining milk traits. The most important metabolites for each milk trait which coincided in both prediction methods are listed in Table 4.6. Comparing the detected important metabolites via RF and PLS (Table 4.6) to metabolites which show significantly large correlation values for the investigated milk traits (Figure 4.7) using the univariate evaluation resulted in good congruence.

Further, the important metabolites are presented separately for each prediction method and also with respect to their importance for the investigated milk traits in Appendix B.4 (on page 157). Comparing both statistical learning methods, more important metabolites were observed with PLS than RF for most milk traits. For casein content and protein content, we found almost the same set of important metabolites with both prediction methods. In total, we found 14 important metabolites which were detected by both

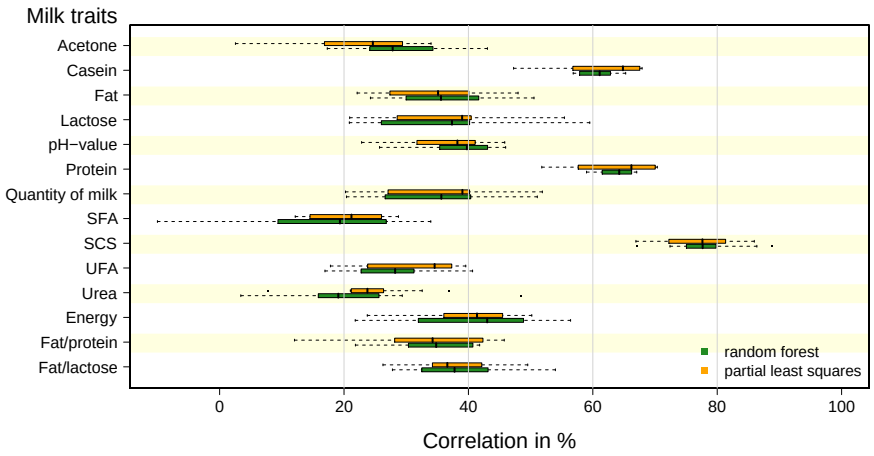


Figure 4.9: Boxplots of observed precisions for the prediction of milk traits (ρ_{milk}; Eq. 4.13) from metabolite profiles using RF and PLS based on a 10-fold cross-validation.

Table 4.6: For each milk trait, the observed important metabolites are listed in alphabetical order. The important metabolites exceeded the 90% quantile for both prediction methods (RF and PLS) in all 10-fold cross-validation runs.

Milk trait	Important metabolites
Acetone (%)	Ethanolaminephosphate; Glucaric acid-1,4-lacton; Orotic acid
Casein (%)	2-Piperidinecarboxylic acid; Adipic acid, 2-amino-; Alanine; Arabitol; Asparagine; Aspartic acid; Butanoic acid, 2-amino-; Cinnamic acid, 3,4,5-trimethoxy-, trans-; Glycerol-3-phosphate; myo-Inositol-1-phosphate; Phosphoenolpyruvic acid; Pyroglutamic acid; Spermidine; Thiazole, 4-methyl-5-hydroxyethyl-
Fat (%)	1,3-Dihydroxyaceton; Arabitol; Aspartic acid; Galactitol; Glucaric acid-1,4-lactone; myo-Inositol-1-phosphate; Pyroglutamic acid
Lactose (%)	1,3-Dihydroxyaceton; Glucaric acid-1,4-lactone; Leucine; Methionine; Phenylalanine; Tyrosine
pH value	Alanine, beta-; Glycerol-2-phosphate; Glycerol-3-phosphate; Glycine
Protein (%)	2-Piperidinecarboxylic acid; Adipic acid, 2-amino-; Arabitol; Asparagine; Aspartic acid; Butanoic acid, 2-amino-; Cinnamic acid, Glyceric acid-3-phosphate; Glycerol-3-phosphate; myo-Inositol-1-phosphate; 3,4,5-trimethoxy-, trans-; Phosphoenolpyruvic acid; Pyroglutamic acid; Spermidine; Thiazole, 4-methyl-5-hydroxyethyl-
Quantity of milk (kg)	Arabitol; Butanoic acid, 2-amino-; Butanoic acid, 4-methylthio-2-oxo-; 2-Piperidinecarboxylic acid
SFA	1,3-Dihydroxyaceton; Glycerol
SCS	1,3-Dihydroxyaceton; Butanoic acid, 2-hydroxy-; Lactic acid, DL; Leucine; Methionine; Phenylalanine; Tryptophan; Tyrosine; Uracil
UFA	Galactitol; Serine, cyclo-
Urea (%)	Adipic acid, 2-amino-; Aspartic acid
Energy (MJ/kg)	1,3-Dihydroxyaceton; Arabitol; Aspartic acid; myo-Inositol-1-phosphate; Pyroglutamic acid
Fat/protein	1,3-Dihydroxyaceton; Butanoic acid, 3-hydroxy-; Galactitol; Glycerol-2-phosphate; Glycerol-3-phosphate; Glucaric acid-1,4-lactone; Sedoheptulose, 2,7-anhydro-, beta-
Fat/lactose	1,3-Dihydroxyaceton; Arabitol; Galactitol; Glucaric acid-1,4-lactone; Pyruvic acid; Pyroglutamic acid

prediction methods for casein content and protein content. In addition, 13 important metabolites also coincided between casein content and protein content, e.g., arabitol, 2-amino adipic acid, pyroglutamic acid. Urea content is an example of disagreement between RF and PLS, for which five metabolites were identified as important by each method, but only aspartic acid and 2-amino-adipic acid coincided between them. In the following, we only consider metabolites detected as important by both prediction methods. For SCS, the most important metabolites were uracil and lactic acid. A positive correlation existed between uracil and SCS ($\rho_{t1t2} = 0.70$), between lactic acid and SCS ($\rho_{t1t2} = 0.58$), and between both these metabolites ($\rho_{t1t2} = 0.62$). Comparing important metabolites for the prediction of lactose and SCS, five important metabolites coincided, i.e., 1,3-dihydroxyaceton, leucine, methionine, phenylalanine and tyrosine. Between SCS and lactose content a negative correlation ($\rho_{t1t2} = -0.28$) was observed, which was also mirrored in the observed correlation structure between metabolites and milk traits (see also Appendix B.1 on page 143). In detail, SCS (lactose content) showed the following correlation values for 1,3-dihydroxyaceton = 0.37 (−0.21), leucine = 0.41 (−0.20), methionine = 0.43 (−0.25), phenylalanine = 0.43 (−0.26) and tyrosine = 0.49 (−0.27). The correlation structures of these metabolites were clustered together (cf. Figure 4.7 part 3). Furthermore, for fat content, energy, and fat:lactose the same three important metabolites were found; arabitol, 1,3-dihydroxyaceton and pyroglutamic acid. We observed that most of the metabolites show similar correlation values regarding these milk traits (cf. Figure 4.7, for further information see Appendix B.1). In this context, we found the best coinciding correlation values for casein content and protein content. Additionally, 1,3-dihydroxyaceton was observed as important for fat content, energy, fat:protein, fat:lactose and SFA. Finally, we did not observe any coinciding important metabolites for fat content, UFA and SFA.

4.3.3 Comparison of the metabolite approach to other approaches

In this section, the results of our proposed metabolite approach versus classical approach, reduced classical approach and QTL approach are presented (cf. Section 4.2.4 on page 83).

4.3.3.1 Determining milk metabolites important for milk traits

Two regression methods (RF and PLS) were applied to determine metabolites important for the investigated milk traits. The observed mean prediction precisions were similar for both regression methods, e.g., $\rho_{milk} = 0.63$ RF and $\rho_{milk} = 0.64$ PLS for protein content (cf. Section 4.3.2.3). The results were based on the inner 10-fold cross-validation. Protein content showed the highest mean prediction precision, whereas the observed mean prediction precisions for fat content and pH value were about 0.35 for both regression methods. The observed important metabolites and their frequency over the 10 inner cross-validation sets are listed in Table 4.7. For fat content, 11 different metabolites

were found to be important, e.g., 1-3-dihydroxyaceton, aspartic acid and galactitol. Six metabolites were found in each inner cross-validation run. In total, ten different important metabolites were found for pH value, of which only glycerol-2-phosphate and glycine were found to be in each inner cross-validation run. For protein content, 16 metabolites were detected as important, and 11 of them were observed in all inner cross-validation runs. Arabitol, aspartic acid and pyroglutamic acid were important for both fat content as well as protein content and they were observed in all inner cross-validation runs. These findings are mainly in congruence with the important detected metabolites in Section 4.3.2.3 where the outer cross-validation was used (cf. Table 4.6).

4.3.3.2 Selecting important SNPs via SVS

The average number of important SNPs for each important metabolite is listed in Table 4.7. In Table 4.8 the average number of important SNPs selected by SVS is listed for the metabolite approach and the reduced classical approach. In general, the average number of important SNPs was larger for the metabolite approach than for the classical approach; at least 42.5% more SNPs were selected. Table 4.8 also presents the number of SNPs in known QTL regions, termed QTL-SNPs, for fat content and protein content. For the QTL approach, at least 12 times as many SNPs were declared important as in the reduced classical approach or metabolite approach. In most cases, the average number of important SNPs was clearly smaller for important metabolites compared to milk traits (reduced classical approach; cf. with Table 4.7). Moreover we observed that the DGAT1-SNP was detected for all three investigated milk traits. Hence it was evaluated how often the DGAT1-SNP was detected over all inner cross-validation sets for important metabolites. The DGAT1-SNP was identified for the following metabolites in all inner cross-validation sets: arabitol, aspartic acid, and pyroglutamic acid, for fat content and protein content. Additionally, the DGAT1-SNP had an impact on 2-amino-butanoic acid and asparagine when studying protein content. For pH value, the DGAT1-SNP was identified nine times based on the metabolite glycine.

4.3.3.3 Enrichment analysis of important SNP subsets with respect to known QTL

For fat content and protein content, it was investigated if sets of metabolite SNPs or reduced SNPs were enriched in the set of QTL-SNPs for all 10 cross-validation sets. In Table 4.9, the observed P-values as well as the number of expected and observed important SNPs located in QTL-SNPs are listed. For both milk traits investigated, the observed P-values were not significant on the significance level $\alpha = 0.05$, except in one case for the reduced classical approach (e.g., P-values were in the range of $[0.048; 0.930]$ for fat content). For the metabolite approach, however, the observed P-values were small and in almost all cases significant (e.g., P-values were in the range of $[0.001; 0.032]$ for fat content). The important SNP positions for each milk trait which were selected in

Table 4.7: Information about important milk metabolites.

Milk trait	Milk metabolite	Frequency in 10 cross-validation runs	Average number of important SNPs
Fat (%)	1,3-Dihydroxyaceton	10	5.30
	Arabitol	10	16.90
	Aspartic acid	10	29.00
	Butanoic acid, 4-amino-	1	4.00
	Galactitol	10	18.10
	Glucaric acid-1,4-lactone	10	7.40
	Muramic acid, N-acetyl-	1	6.00
	myo-Inositol-1-phosphate	8	6.88
	Pyroglutamic acid	10	41.50
	Pyruvic acid	4	10.75
	Sedoheptulose, 2,7-anhydro-, beta	1	4.00
pH value	Alanine, beta-	8	10.00
	Arabitol	3	18.00
	Glutaric acid, 2-hydroxy-	4	33.25
	Glycerol-2-phosphate	10	25.60
	Glycerol-3-phosphate	7	53.57
	Glycine	10	20.60
	Phenylalanine	1	8.00
	Threonic acid	1	4.00
	Tryptophan	1	10.00
	Tyrosine	1	15.00
Protein (%)	2-Piperidinecarboxylic acid	10	23.50
	Adipic acid, 2-amino-	10	24.60
	Alanine	3	9.00
	Arabitol	10	16.30
	Asparagine	10	12.30
	Aspartic acid	10	28.40
	Butanoic acid, 2-amino-	10	29.10
	Cinnamic acid, 3,4,5-trimethoxy-, trans-	10	4.90
	Glyceric acid-3-phosphate	4	23.50
	Glycerol-2-phosphate	1	29.00
	Glycerol-3-phosphate	10	55.10
	myo-Inositol-1-phosphate	10	6.50
	Phosphoenolpyruvic acid	7	36.29
	Pyroglutamic acid	10	41.80
	Spermidine	10	11.20
	Thiazole, 4-methyl-5-hydroxyethyl-	7	11.43

Table 4.8: The average number of selected important SNPs.

Approach	Fat (%)	Protein (%)	pH value
Reduced classical approach	30	88	80
Metabolite approach	129	302	114
QTL approach	3,034	3,593	-

more than seven cross-validation runs with the metabolite approach are presented in Appendix B.6 (on page 162). There it is also marked if a SNP position lies in a known QTL. The SNP positions were specified, and aside from important SNPs in known QTL, further SNP positions were several times detected, indicating their importance for the investigated milk trait.

4.3.3.4 Comparison of prediction results obtained by different approaches using SVS

For all investigated milk traits, boxplots of the observed prediction precisions for each SNP subset approach are presented in Figure 4.10. A significant difference between two approaches regarding the observed prediction precisions is marked with a black dashed line ($\alpha = 0.05$); the corresponding observed P-value is also given. For fat content (Figure 4.10 A), the reduced classical approach ($\rho_{svs}^{milk} = 0.221$) was surpassed by the other three approaches, but no significant differences between the classical approach ($\rho_{svs}^{milk} = 0.299$) and the metabolite approach ($\rho_{svs}^{milk} = 0.290$) and QTL approach ($\rho_{svs}^{milk} = 0.293$) were observed. Compared to the classical approach and the QTL approach, less than 1% of the total amount of (40,317) SNPs were used for the prediction via the metabolite approach. Further, the highest single prediction precision of $\rho_{svs}^{milk} = 0.450$ was observed for the metabolite approach, whereas for the classical approach the highest prediction precision was $\rho_{svs}^{milk} = 0.377$ and $\rho_{svs}^{milk} = 0.430$ for the QTL approach.
For protein content (Figure 4.10 B), the classical approach ($\rho_{svs}^{milk} = 0.237$) outperformed all three other approaches in terms of prediction precision, but no significant difference between the reduced classical approach ($\rho_{svs}^{milk} = 0.147$) and the metabolite approach ($\rho_{svs}^{milk} = 0.126$) or QTL approach ($\rho_{svs}^{milk} = 0.188$) was observed. For pH value (Figure 4.10 C), the observed P-value for the comparison of the classical approach and the metabolite approach is 0.049 which is close to the bound of $\alpha = 0.05$, whereas between the classical approach and the reduced classical approach a clearly significant difference was observed (P-value = 0.002) regarding prediction precisions.
To validate that the mean prediction precisions observed for the metabolite SNP subsets and the reduced SNP subsets were significantly different compared to those of random subsets, we implemented a resampling analysis (cf. Section 4.2.4.2 on page 84). For fat content, a significant difference regarding prediction precision was observed for both

Table 4.9: The P-values from rating the important metabolites for the reduced classical approach and the metabolite approach for each of the 10 cross-validation runs ($\alpha = 0.05$).

	Reduced classical approach			Metabolite approach		
Trait (%)	P-value	Expected	Observed	P-value	Expected	Observed
Fat	0.737	2.56	2	0.010*	7.75	15
	0.588	3.01	3	0.001*	7.30	17
	0.930	2.56	1	0.032*	8.13	14
	0.048*	2.03	5	0.001*	9.26	20
	0.897	2.18	1	0.005*	7.90	16
	0.395	2.26	3	0.003*	9.56	19
	0.904	2.26	1	0.006*	9.48	18
	0.395	2.26	3	0.014*	6.62	13
	0.613	2.03	2	0.008*	8.28	16
	0.807	1.58	1	0.001*	6.77	16
Protein	0.270	7.04	9	0.002*	20.76	35
	0.400	8.91	10	0.073	20.14	27
	0.488	8.56	9	0.017*	23.35	34
	0.921	6.86	4	0.011*	24.24	36
	0.815	7.93	6	0.025*	18.98	28
	0.717	7.04	6	0.004*	27.36	42
	0.690	7.93	7	0.069	28.79	37
	0.268	7.93	10	0.025*	23.97	34
	0.307	7.31	9	0.013*	20.41	31
	0.688	9.00	8	0.050*	18.54	26

approaches, whereas no significant difference occurred for protein content. For pH value, the observed resampling P-value was 0.051 for the metabolite approach and 0.055 for the reduced classical approach. In addition the corresponding residual plots are presented in Appendix B.5 on page 161.

4.4 Discussion

In this chapter, three different investigations to analyze the three system-levels were presented, with the main focus being on the benefit of integrating the metabolome in the GP map. In the first part of this chapter, a purely conceptual comparison of two simulated data sets (conventional approach and SBML approach) and an experimental data set containing information on three milk traits was presented. This comparison represents an alternative way to compare simulated data with experimental data, where the structure of the obtained fastBayesB results were used as the basis. Such a comparison is realized, since, it is of interest to simulate data as realistic as possible which enable methodological development and optimization of estimation methods.

The aim of the second part was to investigate the relationships within and between

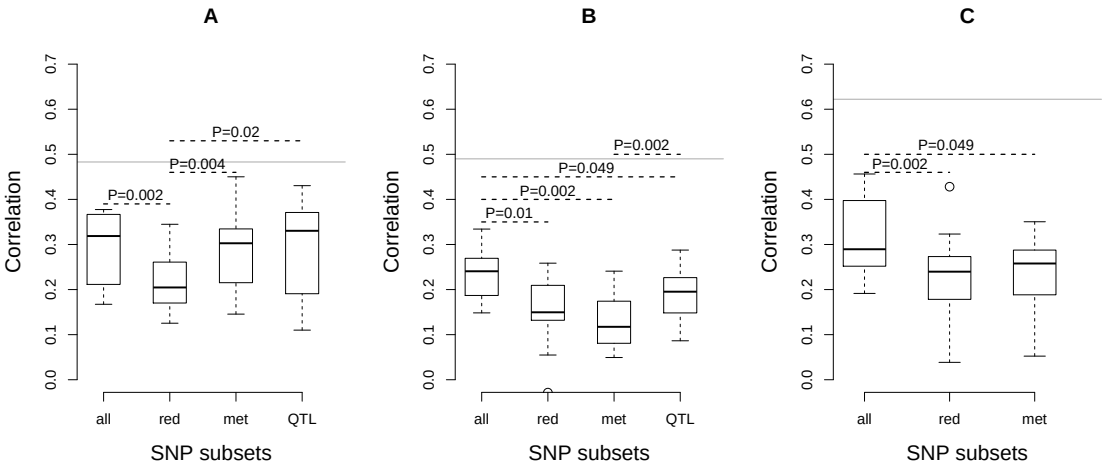


Figure 4.10: Boxplots of the genetic value prediction of ten outer cross-validation runs for the classical approach (all), the reduced classical approach (red), the metabolite approach (met), and the QTL approach (QTL). The following milk traits were investigated: fat content (A), protein content (B) and pH value (C). If two approaches differ significantly ($\alpha = 0.05$), this is marked with a black dashed line and the observed P-value is given. The gray line represents an upper bound for the accuracy of prediction given as the square root of the estimated narrow-sense heritability based on the sire model.

190 milk metabolites and 14 milk traits in more detail. These 190 milk metabolites (representing 10% of the expected milk metabolome) considered only a small part of the metabolic pathways and metabolites important for milk production. The measured milk metabolites represent only a snapshot of the current metabolic state of cows. On the one hand, the impact of influencing factors on milk metabolites as well as milk traits was investigated and reported in greater detail. It could be shown that influencing factors had generally less impact on milk metabolites than on traditional milk traits. On the other hand, significant associations between milk metabolites and milk traits as well as within each level could be characterized, using univariate and multivariate analysis methods. In particular, the application of statistical learning methods (RF and PLS) revealed new relationships between milk metabolites and milk traits.

The third part explored the usefulness of our proposed metabolite approach on experimental data. Three milk traits were analyzed using the metabolite approach and results were compared to three other approaches (classical, reduced classical, QTL approach). To enable fair comparison between the different approaches an invariable analysis design was used. Our analysis show that the prediction precisions using the metabolite approach was more similar to the classical approach than to the reduced classical approach. Moreover a similar mean prediction precision was observed between the classical approach (40,317 SNPs) and the metabolite approach (129 SNPs) for fat content, wherein the metabolite approach required only less than 1% of the total amount of SNPs. An over-representation

analysis revealed that significantly more important SNPs in known QTL regions were selected using the metabolite approach compared to the reduced classical approach.

4.4.1 Experimental data versus simulated data

Our conceptual approach to compare milk traits and different simulated data sets is not meant as direct comparison, as we do not fit any kind of simulation model parameters using our experimental data. The comparison is rather on a conceptual level via comparing the structure of fastBayesB results, because of the unknown underlying number of QTL, i.e., underlying genetic architecture, for the milk traits. The comparison of the composition of the genetic effects for simulated and experimental data offers another perspective to compare experimental data and different alternatives of simulation. Different estimated genetic effect compositions were observed for simulated data sets (Figure 3.6 on page 60) and also for the investigated milk traits (Figure 4.3 on page 89). It was observed that casein content and fat content mainly depend on one major additive effect (which is known from the literature as DGAT1, Grisart et al., 2004), whereas pH value depend on additive as well as dominance genetic effects. Thus, the underlying GP map seems to be more linear in combination with observed σ_a^2 (measure for the degree of linearity) for fat content and protein content than for pH value (cf. Table 4.2 on page 88).
The model used in this analysis seems to be suitable for the experimental data and simulated data, as was judged by visual residual analysis.
In this thesis, only three milk traits of the 11 recorded milk traits were investigated, because the used milk traits had the highest estimated heritability (range=[0.233,0.389]; cf. Table 4.1 on page 82) and were more or less similar to our simulated data sets. The traits were chosen depending on the heritability, since it is known that traits which have a low heritability require a larger sample size than traits with a higher heritability to obtain an acceptable prediction precision in the genetic value prediction and thus to obtain suitable estimates for the genetic variance components (e.g., Daetwyler et al., 2008; Visscher et al., 2008; Daetwyler et al., 2010). Daetwyler et al. (2010) presented a formula which allows an estimation of the expected prediction precision ($E(\rho)$) based on the number of independent chromosomes segments (M_e), number of animals in the parental or training generation (N_p) and the heritability of the trait:

$$E(\rho) = \sqrt{\frac{N_p \cdot h^2}{N_p \cdot h^2 + M_e}}, \tag{4.17}$$

and M_e can be approximated as follows:

$$M_e = 2 \cdot N_{eff} \cdot L, \tag{4.18}$$

where N_{eff} represents the effective population size and L the number of chromosomes. If we assume an $N_{eff} = 100$, $h^2 = 0.3$ and $N_p = 1{,}170$ (based on the cross-validation design

cf. Section 4.2.2 on page 77) for our experimental data, we would obtain $E(\rho) = 0.235$ and $E(\rho) = 0.138$ for $h^2 = 0.1$ in the same setting. The figures show that it is very likely that no meaningful results may be gained for traits with a low heritability. If we want to obtain a prediction value of $E(\rho) \geq 0.5$ then at least 6,665 animals are necessary for $h^2 = 0.3$ or 20,000 for $h^2 = 0.1$, which would surpass this project.

4.4.2 Investigations of relations of milk metabolites and milk traits

In this section the results of the intense investigations of milk metabolites and milk traits are discussed regarding their meaning and possible importance for dairy cattle science.

4.4.2.1 Impact of influencing factors on traits

Studying the impact of influencing factors on traits revealed further insights and associations which will be discussed in the following paragraphs.

Univariate analysis on milk metabolites and milk traits in relation to influencing factors: Influencing factors considered in this thesis were farm, sire effect, (GC-MS batch), day of lactation and test-day, which had significant impact on parts of the 14 milk traits and 190 milk metabolites. Influencing factors were significant for a larger percentage of milk traits compared to metabolites (cf. Table 4.3). Single metabolite profiles seem to be only partly affected by influencing factors in our study, whereas complex traits, such as the investigated milk traits, showed larger dependencies. A possible explanation is that complex traits are composed out of several molecular traits, so that at least some of these are likely to be affected by the considered influencing factor. On the other hand, metabolites as molecular traits are highly interconnected (and correlated) making it difficult to find simple explanations for the observed selective impact of the influencing factors.

Influence of day of lactation on milk metabolites: A lactation curve was expected to exist for some metabolites, such as for the complex milk trait quantity of milk (cf. Figure 1.4 on page 20). Therefore, we used ANOVA to investigate if metabolites showed a significant trend over the lactation days. In general, we found 78 metabolites which had such a trend (cf. Figure 4.4 on page 90 for two examples). As we collected milk samples between the 21st and 120th day of lactation, only a part of the possible lactation curve was observed. The lactation curves for metabolites are probably blurred because all measurements originated from different individuals (1 sample per cow). The cluster analysis (Figure 4.5) also mirrored a clear change of metabolite profiles during the observed part of lactation. Comparing the lactation intervals for metabolites, we found that the first lactation interval showed the most significant differences to other lactation intervals. This finding could be related to the known course of negative EB which can be

observed in the early phase of lactation (cf. Section 1.5.3 on page 18; e.g., Bauman and Currie, 1980; Collard et al., 2000). In Table 4.4, we present those metabolites which had the most significant differences over all pairwise comparisons. These metabolites showed the highest variation over the investigated lactation intervals. A clearer statement of changes and compositions of metabolites could be given if multiple samples from the same cows were taken over the whole lactation period.

Influence of farm on milk metabolites: We detected significant differences in metabolite levels between farms (cf. Figure 4.6 B). It is known from the literature that, among other factors (e.g., keeping conditions, milking, and management), feeding has an important influence on milk composition (Sutton and Morant, 1989; Töpel, 2004). Furthermore odd-chain fatty acids in milk have a high potential to discriminate between different diets, for a recent study see e.g., Cabrita et al. (2003). In this context, we observed two metabolites, i.e., arabitol and 1,3-dihydroxyacetone (Table 4.4), as being important for discriminating between farms in our data. Both examples are carbohydrates, implicating their role as a precursor of fat.
Considering these findings together with the results of our metabolite clustering of farms (Figure 4.6 A), we assume that feeding is the most discriminating aspect in this part of investigation. In addition, some farms show very distinct metabolite profiles, and thus form separate clusters. However, to give a more differentiated and stronger statement, it is necessary to consider different feeding managements and other influencing aspects for the factor farm in greater detail, and to investigate possible relations to the presented results in this thesis.

Univariate versus multivariate analysis of influencing factors: The multivariate analysis (LDA) revealed additional important metabolites compared to univariate analyses for the investigated influencing factors day of lactation and farm (cf. Table 4.5). Important metabolites detected by univariate analysis as well as by multivariate analysis are marked in bold in Table 4.5 for both influencing factors. Multivariate analyses also consider interactions between metabolite profiles and, hence, can detect additional related features compared to the univariate approach (Scheubert et al., 2012). The latter holds also true for the investigations of the relations between milk metabolites and milk traits.

4.4.2.2 Important milk metabolites with multivariate relations to milk traits

To examine the eligibility of milk metabolites as possible biomarker candidates, we analyzed how well traditional milk traits could be predicted based on milk metabolites. To find out which combinations of milk metabolites explain variation in traditional milk traits best, we applied two statistical learning approaches (RF and PLS). Figure 4.9 shows that the milk traits SCS, casein content and protein content could be predicted confidently from metabolite profiles in contrast to the other 11 milk traits. This finding

indicates that the metabolites in the measured data pool include more metabolites relevant for these three milk traits. This conclusion is further supported by the fact that a large number of important metabolites was detected for these milk traits with both prediction methods (cf. Appendix B.4 on page 157). For casein content and protein content, we observed 14 important metabolites, which coincided between both methods (Table 4.6). Furthermore, 13 of these important metabolites also coincided between casein content and protein content. It was expected that nearly the same metabolites would be detected for both milk traits, because casein is a main component of milk protein, constituting over 80% of its mass (Madureira et al., 2007; Töpel, 2004, p. 226). The similarity of both milk traits is also mirrored in Figure 4.7. For casein content and protein content, the amino acids asparagine and aspartic acid, among others, were detected as important. It is known that amino acids are involved in the protein metabolism. It was surprising that only these two amino acids were detected as important for the prediction of both milk traits, taking into account that the measured metabolites include 12 proteinogenic amino acids in total. Alanine was also found to be important for casein content, but only detected as important by one prediction method for protein content. Further investigations revealed that these three metabolites were reproducibly found in contrast to other metabolites. The remaining nine amino acids also showed correlations, but other metabolites showed stronger correlations and, therefore, were included in the presented list instead (cf. Table 4.6 on page 97).

Spermidine, which is a polyamine, was also detected to be important for protein content. It is known that spermidine is involved in the production of protein content (Sanguansermsri et al., 1974; Bolander Jr. and Topper, 1979; Löser, 2000). However, the correlation detected for spermidine and protein content was rather low ($\rho_{t1t2} = 0.28$), indicating a possible large variation either during the course of lactation and/or between individuals, as proposed by Motyl et al. (1995). Likewise, it is possible that spermidine underlies specific technical measurement difficulties as is known for certain metabolites.

SCC is a measure to monitor udder health and is performed as a specific test for mastitis (Schukken et al., 2003; Koivula et al., 2005). We found uracil as the most important metabolite for predicting SCS, which makes it interesting to investigate this relation further because elevated levels of uracil were already proposed to originate from damaged cells (Bi et al., 2000). It is conceivable that such damage would occur during infection. The second important metabolite in our study was lactic acid. Farr et al. (2002) and Davis et al. (2004) proposed that lactic acid can be used as a biomarker for mastitis at an early stage. Furthermore, we detected the following amino acids as related to SCS: tyrosine, methionine, phenylalanine, leucine and tryptophan. Tryptophan has an impact on T-cell proliferation (Frumento et al., 2002; Denis et al., 2007) and, hence, plays a role in the immune system. Additionally, these amino acids, except tryptophan, were also detected as being important for the prediction of lactose content. Between lactose content and SCS, a significant negative correlation was observed. This observation is in line with

results found in the literature, e.g, Klein et al. (2010); Miglior et al. (2007). It is also known that an increased SCC is associated with an increase in lactose (Harmon, 1994). Corresponding to these findings, correlations between common important metabolites and these traits were negative (cf. first and last column of the heatmap in Figure 4.7 on page 94).

To predict levels of SFA in our data set, 1,3-dihydroxyacetone and glycerol were detected as important. Glycerol is important to build milk fat (long chains) in the mammary glands (Luick, 1961).

4.4.2.3 Benefits and constraints of methods

In general, multiple measurements per individual cow would be desirable for future investigations, to enable the dissection of different sources of variance. Promising candidates for biomarkers, however, should also show high correlations between trait and biomarker even without multiple measurements per individual, as we have found for some cases of the investigated milk traits.

In investigations concerning the influence of factors on metabolite profiles, the metabolite measurements were always corrected for all factors except of the one, for which the influence was explored. We could apply a classical LDA because our experimental data set involved sufficiently more samples than features (metabolites). LDA is also appropriate for unbalanced data (Xue and Titterington, 2008). To reveal groupings of the farms regarding mean metabolite profiles, we applied a classical hierarchical clustering involving average linkage. Clustering methods always come along with the uncertainty regarding the number of clusters. Therefore, it is recommended to apply two kinds of criteria, a relative validity criterion and an external criterion (Vendramin et al., 2009). For the influencing factors investigated in our case, these criteria were silhouette score and Jaccard index. Both criteria concordantly indicated three stable clusters for all 18 farms. The clustering results (Figure 4.6 A) were also concordant with obtained results for all pairwise comparisons (Figure 4.6 B). This means that, depending on the similarity between farms in the cluster structure, we observed an increasing number of significant differences between milk metabolites as the similarity between farms decreased. The same was observed for the influencing factor lactation interval. Based on our results for predicting milk traits from metabolites, we could observe similar sets of important metabolites using two different regression methods (RF and PLS). A general threshold for measurements of milk metabolite importance (in this work the 90% quantile was used) cannot be given, because it depends on the investigated trait and measured part of the metabolome. For example, regarding the milk trait protein content, a large number of metabolites showed a high correlation with this milk trait, whereas for urea content a much smaller number of relevant metabolites was found. To search for probable links to known biological functions in further analyses, we used the intersection of sets of important metabolites of both RF and PLS. In most cases, we were able to relate

information about biological functions of the important metabolites to the respective investigated milk trait (cf. Section 4.4.2.2). One of the detected important metabolites (lactic acid) was already proposed as a potential biomarker in recent literature. Thus, metabolite profiles seem eligible as new molecular traits and can be investigated as candidate biomarkers. In this thesis, GC-MS analyses measure only a part of the milk metabolome because short-chain water-soluble metabolites of the energy metabolism are predominantly detected. Other techniques that also explicitly monitor fatty acids and other fat-soluble metabolites could result in valuable complements.

4.4.3 The metabolite approach compared to three other approaches

The presented metabolite approach allows to select SNPs from important metabolites regarding an investigated milk trait, and it represents a new strategy compared to proposed SNP subset selection strategies found in recent literature, e.g., Habier et al. (2009) and Moser et al. (2010). The metabolite approach enables investigations in two different directions. On the one hand, using metabolite profiles to predict milk traits enables the detection of important milk metabolites for an investigated milk trait (Step 1 "Analysis design") and can be further investigated as discussed in Section 4.4.2.2. On the other hand the important milk metabolites were used to determine associated SNPs (Step 2 "Analysis design") which were involved in the milk trait prediction (Step 3 "Analysis design"). Both steps were also performed for the reduced classical approach. Our findings regarding the metabolite approach and the reduced classical approach are discussed in more detail in the following.

4.4.3.1 Metabolite approach versus reduced classical approach

The genetic architecture may be less complex for a metabolite than for a complex milk trait: Significantly more SNPs located in known QTL (under the restriction that not all QTL are known) were detected using the metabolite approach compared to the reduced classical approach (Table 4.9) for fat content and protein content, which was surprising. A possible reason could be that if the complex milk trait itself is investigated that some of the important genetic effects for this milk trait are overlaid. It is imaginable that such important genetic effects are revealed if a less complex trait (possibly a metabolite), which is highly associated with the investigated milk trait, is studied. Moreover it was observed that in most cases, the investigated milk metabolites showed a significantly smaller number of selected important SNPs (Table 4.7) compared to the complex milk traits (Table 4.8), which indicates a less complex underlying genetic architecture.

An SNP with an important impact on a milk trait may also have an impact on at least one important metabolite: We observed, that DGAT1-SNP was detected as important in all cross-validations runs for all three investigated milk traits using the

reduced classical approach. The investigations of protein content and fat content using the metabolite approach revealed that DGAT1-SNP had an impact on at least one metabolite in each cross-validation run. In contrast, DGAT1-SNP was nine times selected for the metabolite glycine via the metabolite approach (Table 4.7). In general, it is not surprising that the DGAT1-SNP position was not detected in all cross-validation runs using the metabolite approach, due to the smaller number of important metabolites for pH value (Table 4.7), but it might be observed if more relevant metabolites are measured for pH value. These observations support our expectation that an SNP with a significant genetic effect on a milk trait also shows a significant genetic effect on at least one of the important metabolites. In this context, I would like to mention that DGAT1 represents a special case because it has a known relevant genetic effect on milk fat and milk protein and its position is validated (cf. Section 1.5.2 on page 17; Weller and Ron, 2011). Hence, this investigations can be seen as a small first evidence, which needs further validation. Our results also indicates that DGAT1 has also an impact on pH value. In an additional analysis, it was quantified how important the DGAT1-SNP position as well as a defined DGAT1-region for each milk trait is. The analysis results are presented in Appendix B.7 on page 164.

Rating important SNP subsets in respect to their role for the investigated milk trait: An over-representation analysis was realized to investigate the relevance of important SNPs selected by the metabolite approach and the reduced classical approach. Significantly more selected important SNPs located in known QTL were detected using the metabolite approach, showing the relevance of most of the selected important SNPs for the investigated milk traits (cf. Table 4.9 on page 102). This possibly indicates that the other selected important SNPs (cf. Appendix B.6 on page 162) could be relevant or are possibly located in unknown QTL for the investigated milk traits which need further analysis. The latter holds true for selected SNPs using the reduced classical approach. The analysis of relevance comes clearly with the restriction that not all QTL are known for a milk trait.

A resampling analysis was applied to quantify the prediction ability of the obtained SNP-subsets of the metabolite approach and the reduced classical approach. The resampling results confirms indirectly that the obtained SNP subsets for the metabolite approach as well as the reduced classical approach are important for fat content and pH value. Findings also indicate that the SNP subsets selected by the metabolite approach are more suitable for the prediction of the investigated milk traits than using the reduced classical approach (cf. Figure 4.10) for fat content and pH value, which needs further analyses and also other milk traits should be compared. In this context, the SNP-subset returned by the metabolite approach might be superior compared the ones returned by reduced classical approach, if more relevant metabolites are measured and analyzed for the investigated trait. Recall that only a part of the milk metabolome (10%) was

measured and investigated.
The resampling approach can also be seen as an indirect measure (indicator) to study the underlying genetic architecture of traits, since for protein content it was observed that no SNP-subset has suitable prediction ability. The latter indicates that protein content probably depends on more loci than fat content and pH value, which needs further investigations.

In the following the obtained prediction precisions (Step 3 "Analysis design") using the metabolite, classical, reduced classical and QTL approach are discussed.

4.4.3.2 SNP subsets in respect to milk trait prediction

For fat content, the observed mean prediction precision was significantly higher for the metabolite approach ($\rho_{svs}^{milk} = 0.290$) than for the reduced classical approach ($\rho_{svs}^{milk} = 0.221$; cf. Figure 4.10). In this case, no significant difference was observed between the classical approach and the metabolite approach, however the metabolite approach required less than 1% of the total amount of 40,317 SNPs. For pH value, the difference regarding the observed prediction precisions between the classical approach and the metabolite approach was very small. In this context, it is expected that no significant difference will be observed if for example a more suitable part of the metabolome is measured for pH value. For protein content it seems difficult to find a suitable SNP subset to obtain an appropriate prediction precision. The reason for this result might be the underlying genetic architecture of this milk trait, since protein content probably depends on many QTL. This assumption is supported by the finding of the resampling analysis, which yielded no significant difference between the metabolite approach and the reduced classical approach with regard to prediction precision.
In general, the QTL approach has two disadvantages:

- First, not all QTL for a trait are known.
- Second, most of the QTL regions comprise a long segment of the corresponding chromosome (Appendix B.2, p. 151). Some of the selected SNPs in these regions are not necessarily important for the investigated milk trait (cf. Section 1.5.2 on page 17), because most QTL regions have a QTL peak location, which is the position with the highest or lowest value depending on the used test statistic. A higher prediction precision might be observed if only the peak locus is considered instead of the whole QTL region, or if different window widths based on the QTL peaks are used for the genetic value prediction.

In general, we were not able to improve the genetic value prediction in the sense of a higher prediction precision, but further interesting relations could be revealed. In this context, it is expected that more than 2,000 metabolites exist in cow's milk (Töpel, 2004).

In this thesis, about 10% of them were analyzed, originating primarily from the central carbon and energy metabolism. We suppose that the prediction precision will increase if more relevant metabolites are measured for the investigated milk trait, which will become possible in the near future as GC-MS databases increase, and with them the possibility to correctly annotate GC-MS profiles.

4.4.3.3 Benefits and constraints of methods

To determine important metabolites for an investigated milk trait the regression methods RF and PLS were applied. Both regression methods were selected based on our findings in Section 4.3.2.3 on page 92 (discussed in Section 4.4.2.3 on page 108) where the same settings were used and reliable results were obtained.
The important SNP subsets for the metabolite approach and the reduced classical approach were analyzed in various tests to determine their relevance and significance for the investigated milk trait more precisely. For this, other data in the form of known QTL were used to enable a first confirmation of some of the important SNPs for the respective investigated milk trait. A resampling approach was realized to quantify the significance of the observed prediction ability for the metabolite approach as well as for the reduced classical approach. In addition, the 95% quantile of the observed ρ_R values (termed ρ_{R95}) was considered to be a suitable measure of the prediction precision which should be obtained at least for the corresponding observed prediction precision (ρ). Furthermore, important SNP subsets of the metabolite approach as well as of the reduced classical approach were tested for their relevance regarding the investigated milk trait using an over-representation analysis. This approach comes truly under the restriction that not all QTL are known for the milk trait, but it can be used as a first indicator.
In the recent literature (e.g., Daetwyler et al. (2008); Visscher et al. (2008)), it is often mentioned, that traits with a low heritability require a larger sample size than traits with moderate or high heritability to obtain an acceptable prediction precision. This implies that more false-positive SNPs would be detected when SNPs are selected for traits with low heritability if the sample size is not adequate (as discussed earlier in cf. Section 4.4.1 on page 104). It is also possible to determine the heritability for each metabolite (i.e., new trait) in order to get an approximation of the expected prediction precision using Eq. 4.17 on page 104. The proposed resampling approach (see above) can also be used to prove the quality of the observed prediction precision for each metabolite. This would allow a deeper insight into the genetic architecture of a metabolite. This kind of information could be also used to improve our proposed metabolite approach, i.e., metabolites could be excluded from analyses which are not eligible for the genetic value prediction. In this context, the ranges of the estimated narrow-sense heritability for our important metabolites were observed in the interval of $[0.076; 0.368]$ for fat content, $[0.110; 0.441]$ for protein content, and $[0.032; 0.492]$ for pH value. The $\hat{h}^2$ values were taken from Wittenburg et al. (2013), where it was found that the observed prediction precision

was lower compared to the expected prediction precision using Eq. 4.17. However, our findings show that, even if the heritability of the metabolite was not taken into account, an appropriate mean prediction precision was achieved for fat content (e.g., $\rho = 0.29$ and $\rho_{R95} = 0.23$ for the metabolite approach) but not for protein content (e.g., $\rho = 0.13$ and $\rho_{R95} = 0.21$ for the metabolite approach). In general, further investigations are necessary to give a recommendation when traits should be excluded due to low heritability.

Three milk traits were chosen based on the following reasons. On the one hand, not for all recorded milk traits (in fact only few) QTL positions can be found in QTL databases. This was, however, a requirement to enable rating of the detected important SNPs. On the other hand to allow a more or less precise detection of SNP (cf. Section 4.4.1 on page 104) we decided to use milk traits which had in our case the highest estimated narrow-sense heritability. The latter was the reason to choose pH value (had highest $\hat{h}^2$), although probably it is not a commonly investigated milk trait in dairy cattle science.

In this thesis, each analysis step of the proposed metabolite approach was evaluated separately, based on the observed results of the previous step. We suppose that an embedded approach optimizing our three step approach in a one-step cross-validation design could be superior. Also, conceivable alternatives would be to use other data from the metabolome or genome level to optimize filter criteria or to use such information for weighting SNPs.

Finally, it is recommended to evaluate our approach for the inclusion of non-additive effects (Lee et al., 2008; Toro and Varona, 2010). In this work, only the additive genetic effects were considered, due to the large number of analyses. On a 2.93 GHz multi-user system, a Gibbs-sampler round using a purely additive model needs approximately 18 hours and an additive-dominance model needs approximately about 28 hours (personal communication with Dr. D. Wittenburg).

4.4.4 Additional aspects of modeling

4.4.4.1 Correction for known influencing factors

As mentioned earlier (cf. Section 1.5.3 and on page 18) it is necessary to correct for known influencing factors for milk traits. Based on this knowledge the milk metabolites were corrected. To enable the correction a randomized design for measuring the metabolite spectra was created (cf. Chapter 2). Here, we decided to correct for the influencing factors similar to a test-day model (Ptak and Schaeffer, 1993) for both, milk metabolites and milk traits. The large amount of data (more than 1,300 cows) allows us to consider the influencing factors of farm and test-day as cross-classified. As all cows of a farm were fed with the same feed, an effect of diet was also considered by the effect of farm x test-day. The linear and quadratic regression on lactation day were involved to account for a variable state of composition of milk or to account for metabolic changes between 21st and 120th day of lactation. In general it is expected that outside of our investigated

lactation period a stronger deviation of milk composition as well as changes in milk metabolites might be observed among the animals, which would require a higher order regression.

4.4.4.2 Using the sire model

For the estimation of the heritability several opportunities exist, for example, the implementation of an animal model with consideration of the complete pedigree information or without. The most commonly used model in the field of dairy cattle science is the animal model (Visscher et al., 2008). This model is more accurate in some cases than the sire model, because if the sire model is used the correction is realized only for bulls but not for dams which lead to a bias. The advantage of the sire model is that it needs clearly less equations (Mrode, 2005, p. 52-55), i.e., the covariance matrix does not need to be created, and thus can be estimated without much effort. In our case the sire model was suitable, since the heritability was mainly estimated to obtain an upper bound for the precision of the genetic value prediction for the investigated milk trait. The sire model was also adequate to account for the half-sib structure in the data and allowed a better correction for the fixed effects for the investigations of relationships between milk metabolites and milk traits. In this thesis all analyses regarding genetic value prediction were realized without considering the pedigree information, since phenotypes and marker data are adequate and sufficient for the genetic value prediction (Hayes and Goddard, 2010). In addition, the narrow-sense heritability of our milk traits was also estimated based on an animal model by Dr. D. Wittenburg, including additionally the pedigree information. The estimated heritabilities of the animal model as well as of the sire model were similar.

4.5 Summary

In this chapter a purely conceptual comparison between experimental data and different simulated data was presented to represent another perspective to compare such data regarding to their composition of genetic effect sizes and – finally – with regard to the eligibility of simulated data for methodological development and optimization.
In the second part of this chapter deeper investigations of the relationship between milk metabolites and milk traits were presented as well as the degree of impact of the influencing factors on milk traits and milk metabolites. In general, we could observe that the degree of the impact of the influencing factors was less pronounced for milk metabolites compared to milk traits. Two influencing factors (farm and lactation interval) were investigated more intensely regarding milk metabolites, for which significant differences were detected. Studying the relationships between milk metabolites and milk traits as well as within each level revealed significant associations. Deeper investigations of the relationship between metabolites and milk traits revealed some known biological relations, for example lactic

acid for SCS as well as new relationships, which need further analyses and in particular for their use as possible biomarker candidates for traits of interest.

Finally, in this chapter the proposed metabolite approach was tested and compared to the classical as well as reduced classical approach. In addition, the QTL approach could be applied only for two of the investigated milk traits, since not all QTL positions affecting milk traits are known or can be found in databases. An invariable analysis (double ten fold cross-validation) design was used to enable comparability between the different approaches. In this design it was also considered to account for the known half-sib structure in the experimental data. Our analyses revealed that the metabolite approach resulted in a more similar precision for the genetic value prediction to the classical approach as the reduced classical approach for our analyzed milk traits. In this context a relevant observation is that less than one percent of the total amount of SNPs were necessary for the metabolite approach to obtain a prediction precision similar to the classical approach using all SNPs for fat content. Another interesting observation was that significantly more important SNPs detected via the metabolite approach were located in known QTL than using the reduced classical approach. This fact supports our assumption that metabolites may have a less complex underlying genetic architecture compared to complex milk traits. Finally, a resampling approach was proposed to validate the quality of selected SNPs for a milk trait regarding prediction ability, which revealed further information of the underlying genetic architecture.

5 Conclusions

The objective of this thesis was to investigate if it is possible to improve genetic value prediction by considering additional information about the metabolome level. To address this objective, the corresponding data (genotype, metabolome, phenotype) were simulated as well as experimental data collected. Both kinds of data were analyzed using the newly developed integrative bioinformatics approach.

Conclusions for the experimental data collection:

- Through the entire data collection it is necessary to know which kind of data are complete, missing or erroneous to obtain the desired information complete for a specific number of animals. From our experiences during data collection, we recommend to use a database, which represents a useful and helpful instrument, and to write corresponding analysis scripts, that can then be easily and reproducibly applied for validation during the whole collection time.

- In this thesis it was necessary to include 1,834 animals in the study to arrive at 1,305 complete records. General advice, however, on how much animals should be selected to achieve the desired number of complete animal records cannot be given, since it depends on the duration as well as on the kind of experiment.

- Known influencing factors should be taken into account if milk metabolites are measured via GC-MS as realized in this thesis. This will probably also hold true if another technique for extracting the metabolites is used, to enable an unbiased analysis.

Conclusions for the data simulation:

- The SBML approach is proposed to simulate more realistic data as in the conventional approach, based on a more complex model of the GP map. The SBML approach makes use of a metabolic network as additional level of the GP map.

- As the conventional approach, our proposed SBML approach is also artificial, especially the second step to build genetic values as discussed in Chapter 3. A main difference between both simulation approaches lies in simulating the genetic effects. In the SBML approach, these genetic effects were implicitly simulated by the interactions of the metabolic network model, wherein some enzyme kinetic parameters of the model were varied in dependency of QTL. This approach represents

a more biological realization to simulate genetic effects, whereas genetic effects are simulated in a statistical sense in the conventional approach.

- A deeper investigation of the simulated genetic effects for the SBML approach revealed that some of the genetic effect sizes were negligible, i.e., some simulated QTL are not important for the phenotype because their impact on the metabolic outcome is already very low. However, such QTL positions in real data could be important or become measurable if the network architecture of regulatory interactions changes, for example, in case of a disease.

- Based on the last two statements, it can be recommended to simulate a GP map including a molecular level to explore the importance of the genetic variation on this intermediate level and its transformation through molecular networks.

- A main advantage of the SBML approach is to have three kinds of data from different system-levels. These kinds of data allow various further methodological investigations and optimizations as well as to test different analysis possibilities. In this context, an integrative bioinformatics approach, i.e., the metabolite approach, was proposed to verify on simulated data, if and to which degree it is possible to improve the genetic value prediction when considering the metabolome or just a part of it. Also, these kinds of data can be used for finding suitable measures, for example, finding appropriate weights for SNPs to improve genetic value prediction. Such weights could be developed depending on the importance of simulated metabolites for phenotype prediction as it was realized in this thesis.

- The classical analysis of the GP map using a linear model (considering additive and dominance genetic effects) was realized for data simulated by the conventional approach and by the SBML appraoch. Results revealed that data simulated with the SBML approach show a smaller degree of linearity (σ_a^2), and lower prediction precisions for the different investigated scenarios, compared to the conventional approach. In particular, for the conventional approach the genetic effects (without simulating epistatic genetic effects) were simulated in a specified range and can be explicitly partitioned by the researcher (e.g., into additive and dominance genetic effects). In contrast, the size of the different kinds of genetic effects of the SBML model were unknown.

- Our purely conceptual comparison between simulated and experimental data revealed that similarities can be found regarding genetic architecture, with trait specific details. In general, it will depend on the genetic architecture of the experimental trait (n_{QTL} is unknown), which of the simulation approaches are more appropriate. From this follows that no general statement can be made whether the alternative SBML approach is more realistic than the conventional approach for data simulation.

Conclusions of intense investigations of the relationship between the metabolome and phenotypes using the experimental data:

- Based on our results of the intense investigations on milk metabolites and milk traits also in view of influencing factors, we also recommend to use uni- and multivariate analysis methods to analyze such kinds of data in more detail as usual in other fields. Using multivariate analysis, where also the interactions between metabolites are considered, allows to find further relevant relationships.

- Metabolite profiles proved to be promising new molecular traits eligible to be investigated as candidate biomarkers or to be used in groups of important metabolites (biosignatures). In this context, it is necessary to further elucidate metabolites' physiological role, and also to validate the important metabolites revealed in our work with another data set. Another aspect to be considered is that many of the measured metabolites are used for synthesis of milk components by the alveolar epithelial cells or are involved in the intracellular metabolism. This leads to the question: why were they measured in milk?
  In order to prove the suitability of metabolites as biomarker candidates, it is also necessary to investigate to which degree they change during the course of individual lactations. Taking multiple samples per cow during lactation would help to discriminate technical as well as biological sources of variation.

- As far as, for example, mastitis is concerned, it could be conceived that SCS compared to a traditional biomarker can monitor a different aspect of the disease as a related metabolite (set), as is already known for lactic acid, which specifically shows a relation to the early onset of mastitis. It could also be beneficial to further investigate the correlation structure (e.g., partial correlations or mutual correlations) of the metabolite profiles to reveal possible associations or functional grouping structures. Such structures could be further related to a-priori functional knowledge, for example metabolite pathways or flux modes. Findings from such investigations could further illuminate the functional basis of candidate biosignatures and may help improve learning algorithms exploiting biosignatures which are able to use such a-priori information.

- Based on clustering results for metabolite profiles regarding influencing factor farm in this thesis, it could be conceived that deeper investigations of this factor would allow for a comparison of the different management systems and feeding regimes and help to find out how these are mirrored in the differences of the metabolite profiles, probably revealing other important factors. The resulting consequences could lead to improvements in the field of farming.

Conclusions from implementing the integrative bioinformatics approach (metabolite approach) for the analysis of experimental data as compared to the classical

approach:

- The proposed metabolite approach was used and compared to three different approaches, resulting in various SNP subsets, regarding the prediction precision for three selected milk traits. It was observed that using the metabolite approach led to a similar prediction precision compared to the classical approach, but required less than 1% of the total amount of (40,317) SNPs. In most cases, the metabolite approach performed better than the reduced classical approach.

- The number of selected important SNPs were mostly lower for milk metabolites than for milk traits. This result indicates that most of milk metabolites have a less complex underlying genetic architecture compared to milk traits. In this context it was expected that SNPs with a significant impact on milk traits should also show a genetic impact at least on one milk metabolite, which can also be seen as an additionally indicator for the relevance of the metabolite for the investigated milk trait. In this thesis DGAT1 was used as an example. In this context the relevance of selected SNPs via the metabolite approach and reduced classical approach was also investigated in a very small framework in this thesis. The relevance of some of the selected SNPs for an investigated milk trait was confirmed using enrichment analysis based on known QTL for a milk trait (under the restriction that not all QTL are known), which revealed that significantly more important SNPs were located on known QTL detected by the metabolite approach. Thus it is recommended to further validate the detected important SNP positions regarding their relevance for the investigated milk trait in more detail.

- We recommend to assess the significance of the number of selected important SNPs for the genetic value prediction. In this thesis it was realized using a resampling approach. This offers a further possibility to learn more about the underlying genetic architecture of an investigated trait.

- The success of the metabolite approach regarding prediction precision depends, among other things, on the underlying genetic architecture of the investigated milk trait, and presumably on the measured part of the milk metabolome.

Final conclusions which can be drawn from this thesis, regarding a further intermediate level of the classical GP map: More realistic simulation of all levels of data is recommended and should preferably be based on experimental data collected, especially if new unknown relationships will be investigated or different kinds of data are combined for investigations. It is recommend to apply different analysis techniques to enable insight from different perspectives, which allows discovery of new interesting associations or identification of concordances from different applied methods (univariate and multivariate). All these investigations can lead to a deeper understanding of underlying associations

of the complex system. The proposed metabolite approach in this thesis allows deeper insight into the associations between the different system-levels of the more complex GP map and enables similar genetic value prediction performance as state-of-the-art methods, but based on a significantly smaller subset of SNP markers.

Bibliography

M. Ackermann and K. Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47. 2009.

ADR. *ADR-Handbuch Empfehlungen und Richtlinien*, Arbeitsgemeinschaft Deutscher Rinderzüchter e.V., Bonn, Germany. 2008.

A. K. A. Ali and G. E. Shook. An optimum transformation for somatic cell concentration in milk. *J. Dairy Sci.*, 63(3):487–490. 1980.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410. 1990.

J. M. Álvarez Castro and Ö. Carlborg. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*, 176(2):1151–1167. 2007.

L. Andersson. Genetic dissection of phenotypic diversity in farm animals. *Nat. Rev. Genet.*, 2(2):130–138. 2001.

J.-J. Arranz, W. Coppieters, P. Berzi, N. Cambisano, B. Grisart, L. Karim, F. Marcq, L. Moreau, C. Mezer, J. Riquet, P. Simon, P. Vanmanshoven, D. Wagenaar, and M. Georges. A QTL affecting milk yield and composition maps to bovine chromosome 20: a confirmation. *Anim. Genet.*, 29(2):107–115. 1998.

M. S. Ashwell, Y. Da, C. P. Van Tassell, P. M. VanRaden, R. H. Miller, and C. E. Rexroad Jr. Detection of putative loci affecting milk production and composition, health, and type traits in a United States Holstein population. *J. Dairy Sci.*, 81(12):3309–3314. 1998.

M. S. Ashwell, D. W. Heyen, T. S. Sonstegard, C. P. Van Tassell, Y. Da, P. M. VanRaden, M. Ron, J. I. Weller, and H. A. Lewin. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. *J. Dairy Sci.*, 87(2):468–475. 2004.

A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, J. Woodcock, and S. L. Zeger. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.*, 69(3):89–95. 2001.

J. E. Bailey. Complex biology with no parameters. *Nat. Biotechnol.*, 19(6):503–504. 2001.

D. E. Bauman and W. B. Currie. Partitioning of nutrients during pregnancy and lactation: a review of mechanisms involving homeostasis and homeorhesis. *J. Dairy Sci.*, 63(9):1514–1529. 1980.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B (Methodological)*, 57(1):289–300. 1995.

J. Bennewitz and T. H. E. Meuwissen. The distribution of QTL additive and dominance effects in porcine F2 crosses. *J. Anim. Breed. Genet.*, 127(3):171–179. 2010.

J. Bennewitz, N. Reinsch, J. Szyda, F. Reinhardt, C. Kühn, M. Schwerin, G. Erhardt, C. Weimann, and E. Kalm. 2003. Marker assisted selection in German Holstein dairy cattle breeding: outline of the program and marker assisted breeding value estimation. In *Book of Abstracts of the 62nd Annual Meeting of the EAAP*. Wageningen Academic Publishers, The Netherlands, p. 5. ISSN 1382-6077.

D. Bi, L. W. Anderson, J. Shapiro, A. Shapiro, J. L. Grem, and C. H. Takimoto. Measurement of plasma uracil using gas chromatography-mass spectrometry in normal individuals and in patients receiving inhibitors of dihydropyrimidine dehydrogenase. *J. Chrom. B Biomed. Sci. Appl.*, 738(2):249–258. 2000.

F. F. Bolander Jr. and Y. J. Topper. Relationships between spermidine, glucocorticoid and milk proteins in different mammalian species. *Biochem. Biophys. Res. Commun.*, 90(4):1131–1135. 1979.

L. Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32. 2001.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81(5):1084–97. 2007.

A. R. J. Cabrita, A. J. M. Fonseca, R. J. Dewhurst, and E. Gomes. Nitrogen supplementation of corn silages. 2. assessing rumen function using fatty acid profiles of bovine milk. *J. Dairy Sci.*, 86(12):4020–4032. 2003.

M. P. L. Calus, T. H. E. Meuwissen, A. P. W. De Roos, and R. F. Veerkamp. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, 178(1):553–561. 2008.

M. P. L. Calus and R. F. Veerkamp. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J. Anim. Breed. Genet.*, 124(6):362–3688. 2007.

Ö. Carlborg, L. Jacobsson, P. Åhgren, P. Siegel, and L. Andersson. Epistasis and the release of genetic variation during long-term selection. *Nat. Genet.*, 38(4):418–420. 2006.

B. Charlesworth. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.*, 10(3):195–205. 2009.

H. Y. Chen, Q. Zhang, C. C. Yin, C. K. Wang, W. J. Gong, and G. Mei. Detection of quantitative trait loci affecting milk production traits on bovine chromosome 6 in a Chinese Holstein population by the daughter design. *J. Dairy Sci.*, 89(2):782–790. 2006.

H. Choi and N. Pavelka. When one and one gives more than two: challenges and opportunities of integrative omics. *Front. Genet.*, 2:105. 2012.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74(368):829–836. 1979.

M. Cohen-Zinder, E. Seroussi, D. M. Larkin, J. J. Loor, A. Everts-van der Wind, J.-H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, H. A. Lewin, J. I. Weller, and M. Ron. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res.*, 15(7):936–944. 2005.

B. C. Y. Collard, M. Z. Z. Jahufer, J. B. Brouwer, and E. C. K. Pang. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica*, 142(1):169–196. 2005.

B. L. Collard, P. J. Boettcher, J. C. M. Dekkers, D. Petitclerc, and L. R. Schaeffer. Relationships between energy balance and health traits of dairy cattle in early lactation. *J. Dairy Sci.*, 83(11):2683–2690. 2000.

H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, 11(20):2463–2468. 2002.

Á. Cuadros-Inostroza, C. Caldana, H. Redestig, M. Kusano, J. Lisec, H. Peña-Cortés, L. Willmitzer, and M. A. Hannah. TargetSearch - a Bioconductor package for the efficient pre-processing of GC-MS metabolite profiling data. *BMC Bioinformatics*, 10:428. 2009.

H. D. Daetwyler, R. Pong-Wong, B. Villanueva, and J. A. Woolliams. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185(3):1021–1031. 2010.

H. D. Daetwyler, B. Villanueva, and J. A. Woolliams. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE*, 3(10):e3395. 2008.

S. R. Davis, V. C. Farr, C. G. Prosser, G. D. Nicholas, S.-A. Turner, J. Lee, and A. L. Hart. Milk L-lactate concentration is increased during mastitis. *J. Dairy Res.*, 71(2):175–181. 2004.

H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, 9(1):67–103. 2002.

S. Dejean, I. Gonzalez, K.-A. Le Cao, and P. Monget, mixOmics: Omics Data Integration Project. 2011. URL `http://CRAN.R-project.org/package=mixOmics`, R package version 2.9-6.

J. C. M. Dekkers. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.*, 82(13). 2004.

J. C. M. Dekkers. Application of genomics tools to animal breeding. *Curr. Genomics*, 13(3):207–212. 2012.

J. C. M. Dekkers and J. P. Gibson. Applying breeding objectives to dairy cattle improvement. *J. Dairy Sci.*, 81(Suppl 2):19–35. 1998.

M. Denis, D. N. Wedlock, A. R. McCarthy, N. A. Parlane, P. J. Cockle, H. M. Vordermeier, R. G. Hewinson, and B. M. Buddle. Enhancement of the sensitivity of the whole-blood gamma interferon assay for diagnosis of Mycobacterium bovis infections in cattle. *Clin. Vaccine Immunol.*, 14(11):1483–1489. 2007.

I. R. Dohoo and A. H. Meek. Somatic cell counts in bovine milk. *Can. Vet. J.*, 23(4):119–125. 1982.

S. Drăghici, P. Khatri, R. P. Martins, G. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104. 2003.

J. S. Edwards, M. Covert, and B. Palsson. Metabolic modelling of microbes: the flux-balance approach. *Environ. Microbiol.*, 4(3):133–140. 2002.

F. Enjalbert, M. C. Nicot, C. Bayourthe, and R. Moncoulon. Ketone bodies in milk and blood of dairy cows: relationship between concentrations and utilization for detection of subclinical ketosis. *J. Dairy Sci.*, 84(3):583–589. 2001.

Ensembl. 2008. URL `http://www.ensembl.org/Bos_taurus`, accessed November 2008.

D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics*, Longmans Green, Harlow, Essex, UK. 4 ed. 1996.

V. C. Farr, C. G. Prosser, G. D. Nicholas, J. Lee, A. L. Hart, and S. R. Davis. Increased milk lactic acid concentration is an early indicator of mastitis. *Proceedings of the New Zealand Society of Animal Production*, 62:22–23. 2002.

R. L. Fernando and M. Grossman. Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.*, 21(4):467–477. 1989.

O. Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48(1-2):155–171. 2002.

V. Fievez, B. Vlaeminck, M. S. Dhanoa, and R. J. Dewhurst. Use of principal component analysis to investigate the origin of heptadecenoic and conjugated linoleic acids in milk. *J. Dairy Sci.*, 86(12):4047–4053. 2003.

R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinb.*, 52:399–433. 1918.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7(2):179–188. 1936.

G. Frumento, R. Rotondo, M. Tonetti, G. Damonte, U. Benatti, and G. B. Ferrara. Tryptophan-derived catabolites are responsible for inhibition of t and natural killer cell proliferation induced by indoleamine 2,3-dioxygenase. *J. Exp. Med.*, 196(4):459–468. 2002.

H. Gao, M. Fang, J. Liu, and Q. Zhang. Bayesian shrinkage mapping for multiple QTL in half-sib families. *Heredity (Edinb)*, 103(5):368–376. 2009.

H. Ge, A. J. M. Walhout, and M. Vidal. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, 19(10):551–560. 2003.

T. Geishauser, K. Leslie, J. Tenhag, and A. Bashiri. Evaluation of eight cow-side ketone tests in milk for detection of subclinical ketosis in dairy cows. *J. Dairy Sci.*, 83(2):296–299. 2000.

H. Geldermann. Investigations on inheritance of quantitative characters in animals by gene markers. *Theor. Appl. Genet.*, 46(7):319–330. 1975.

M. E. Goddard and B. J. Hayes. Genomic selection. *J. Anim. Breed. Genet.*, 124(6):323–330. 2007.

P. Good. *Permutation, parametric and bootstrap tests of hypotheses*, 3 ed., Springer-Verlag, New York, USA. Springer Series in Statistics. 2005.

B. Grisart, F. Farnir, L. Karim, N. Cambisano, J.-J. Kim, A. Kvasz, M. Mni, P. Simon, J.-M. Frère, W. Coppieters, and M. Georges. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl. Acad. Sci. USA*, 101(8):2398–2403. 2004.

R. R. Grummer and R. R. Rastani. Review: When should lactating dairy cows reach positive energy balance? *The Professional Animal Scientist*, 19(3):197–203. 2003.

D. Habier, R. L. Fernando, and J. C. M. Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397. 2007.

D. Habier, R. L. Fernando, and J. C. M. Dekkers. Genomic selection using low-density marker panels. *Genetics*, 182(1):343–353. 2009.

D. Habier, R. L. Fernando, K. Kizilkaya, and D. J. Garrick. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12:186. 2011.

J. B. Hagen. The origins of bioinformatics. *Nat. Rev. Genet.*, 1(3):231–236. 2000.

J. B. S. Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.*, 8:299–309. 1919.

R. J. Harmon. Physiology of mastitis and factors affecting somatic cell counts. *J. Dairy Sci.*, 77(7):2103–2112. 1994.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*, 2 ed., Springer, New York, USA. Springer Series in Statistic. 2009.

B. Hayes and M. Goddard. Genome-wide association and genomic selection in animal breeding. *Genome*, 53(11):876–883. 2010.

R. Heinrich and S. Schuster. The modelling of metabolic systems. structure, control and optimality. *Biosystems*, 47(1-2):61–77. 1998.

C. R. Henderson. Estimation of changes in herd environment. *J. Dairy Sci.*, 32:706–711. 1949.

C. Hennig. Cluster-wise assessment of cluster stability. *Comput. Stat. Data An.*, 52(1):258–271. 2007.

C. Hennig, fpc: Flexible procedures for clustering. 2010. URL `http://CRAN.R-project.org/package=fpc`, R package version 2.0-3.

D. W. Heyen, J. I. Weller, M. Ron, M. Band, J. E. Beever, E. Feldmesser, Y. Da, G. R. Wiggans, P. M. VanRaden, and H. A. Lewin. A genome scan for QTL influencing milk production and health traits in dairy cattle. *Physiol. Genomics*, 1(3):165–175. 1999.

W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.*, 4(2):e1000008. 2008.

W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, 38(6):226–231. 1968.

H.-G. Holzhütter. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.*, 271(14):2905–2922. 2004.

Z.-L. Hu, E. R. Fritz, and J. M. Reecy. AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucl. Acids Res.*, 35(suppl 1):D604–D609. 2007.

M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, and the rest of the SBML Forum:, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531. 2003.

T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, 2:343–372. 2001.

Illumina. 2008. URL `http://www.illumina.com/`, accessed Aug 2008.

Illumina. 2012. URL `http://www.illumina.com`, accessed Dec 2012.

K. L. Ingvartsen, R. J. Dewhurst, and N. C. Friggens. On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. *Livest. Prod. Sci.*, 83(2-3):277–308. 2003.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Stat.*, 33(2):730–773. 2005.

P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaud. Sci. Nat.*, 37:547–579. 1901.

D. A. James and S. DebRoy, RMySQL: R interface to the MySQL database. 2006. URL `stat.bell-labs.com/RS-DBI;www.mysql.com;www.omegahat.org`, R package version 0.5-7.

A. R. Joyce and B. Ø. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.*, 7(3):198–210. 2006.

M. Kærn, W. J. Blake, and J. J. Collins. The engineering of gene regulatory networks. *Annu. Rev. Biomed. Eng.*, 5:179–206. 2003.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28(1):27–30. 2000. Accesssed Nov 2008.

S. M. Kappes, J. W. Keele, R. T. Stone, R. A. McGraw, T. S. Sonstegard, T. P. Smith, N. L. Lopez-Corrales, and C. W. Beattie. A second-generation linkage map of the bovine genome. *Genome Res.*, 7(3):235–249. 1997.

M. Kirchgessner. *Tierernährung*, DLG-Verlag, Frankfurt, Germany. 8 ed. 1992.

H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664. 2002.

M. S. Klein, M. F. Almstetter, G. Schlamberger, N. Nürnberger, K. Dettmer, P. J. Oefner, H. H. D. Meyer, S. Wiedemann, and W. Gronwald. Nuclear magnetic resonance and mass spectrometry-based milk metabolomics in dairy cows during early and late lactation. *J. Dairy Sci.*, 93(4):1539–1550. 2010.

M. S. Klein, N. Buttchereit, S. P. Miemczyk, A.-K. Immervoll, C. Louis, S. Wiedemann, W. Junge, G. Thaller, P. J. Oefner, and W. Gronwald. NMR Metabolomic Analysis of Dairy Cows Reveals Milk Glycerophosphocholine to Phosphocholine Ratio as Prognostic Biomarker for Risk of Ketosis. *J. Proteome Res.*, 11(2):1373–1381. 2012.

M. Koivula, E. A. Mäantysaari, E. Negussie, and T. Serenius. Genetic and phenotypic relationships among milk yield and somatic cell count before and after clinical mastitis. *J. Dairy Sci.*, 88(2):827–833. 2005.

S. König, H. Simianer, and A. Willam. Economic evaluation of genomic breeding programs. *J. Dairy Sci.*, 92(1):382–391. 2009.

J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, P. Dörmann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, and D. Steinhauser. Gmd@csb.db: the golm metabolome database. *Bioinformatics*, 21(8):1635–1638. 2005.

D. D. Kosambi. The estimation of map distance from recombination values. *Ann. Eugen.*, 12(1):172–175. 1944.

C. Y. Kramer. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, 12(3):307–310. 1956.

A. Krastanov. Metabolomics - the state of art. *Biotechnol. Biotec. Eq.*, 24(1):1537–1543. 2010.

L. E. B. Kruuk and J. D. Hadfield. How to separate genetic and environmental causes of similarity between relatives. *J. Evol. Biol.*, 20(5):1890–1903. 2007.

C. Kühn, G. Thaller, A. Winter, O. R. P. Bininda-Emonds, B. Kaupe, G. Erhardt, J. Bennewitz, M. Schwerin, and R. Fries. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. *Genetics*, 167(4):1873–1881. 2004.

J. Kučerová, M. S. Lund, P. Sørensen, G. Sahana, B. Guldbrandtsen, V. H. Nielsen, B. Thomsen, and C. Bendixen. Multitrait quantitative trait loci mapping for milk production traits in danish Holstein cattle. *J. Dairy Sci.*, 89(6):2245–2256. 2006.

M. K. Lau, DTK: Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test Adjusted for Unequal Variances and Unequal Sample Sizes. 2011. URL `http://CRAN.R-project.org/package=DTK`, R package version 3.1.

N. Le Novère, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J. L. Snoep, and M. Hucka. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, 34(Suppl 1):D689–D691. 2006.

E. Lecoutre. The R2HTML package. *R News*, 3(3):33–36. 2003. URL `http://cran.r-project.org/doc/Rnews/Rnews_2003-3.pdf`.

S. H. Lee, J. H. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.*, 4(10):e1000231. 2008.

Z. Lei, D. V. Huhman, and L. W. Sumner. Mass spectrometry strategies in metabolomics. *J. Biol. Chem.*, 286(29):25425–25442. 2011.

R. C. Lewontin and K.-I. Kojima. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution*, 14(4):458–472. 1960.

A. Liaw and M. Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22. 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

J. Lisec, N. Schauer, J. Kopka, L. Willmitzer, and A. R. Fernie. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protoc.*, 1(1):387–396. 2006.

B. Liu, A. de la Fuente, and I. Hoeschele. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, 178(3):1763–1776. 2008.

N. Long, D. Gianola, G. J. M. Rosa, K. A. Weigel, A. Kranis, and O. González-Recio. Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genet. Res. Camb.*, 92(3):209–225. 2010.

C. Löser. Polyamines in human and animal milk. *Br. J. Nutr.*, 84(Suppl 1):S55–S58. 2000.

J. R. Luick. Synthesis of milk fat in the bovine mammary gland. *J. Dairy Sci.*, 44(4):652–657. 1961.

N. M. Luscombe, D. Greenbaum, and M. Gerstein. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.*, 40(4):346–358. 2001.

T. F. C. Mackay. The genetic architecture of quantitative traits. *Annu. Rev. Genet.*, 35:303–339. 2001.

M. J. Mackinnon and M. A. J. Georges. Marker-assisted preselection of young dairy sires prior to progeny-testing. *Livest. Prod. Sci.*, 54(3):229–250. 1998.

A. R. Madureira, C. I. Pereira, A. M. P. Gomes, M. E. Pintado, and F. X. Malcata. Bovine whey proteins - overview on their main biological properties. *Food Res. Int.*, 40(10):1197–1211. 2007.

MATLAB. The MathWorks Inc., Natick, Massachusetts. *7.9.0.529 (R2009b)*. 2009.

N. Melzer, S. Jakubowski, S. Hartwig, U. Kesting, S. Wolf, F. Reinhardt, E. Pasman, R. Nürnberg, G., N., and D. Repsilber. 2010a. Design, infrastructure and database structure for a study on predicting of milk phenotypes from genome wide snp markers and metabolite profiles. Abstract ID 0427. In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*. Gesellschaft für Tierzuchtwissenschaften e.V., Leipzig, Germany.

N. Melzer, D. Wittenburg, S. Hartwig, S. Jakubowski, U. Kesting, L. Willmitzer, J. Lisec, N. Reinsch, and D. Repsilber. Investigating associations between milk metabolite profiles and milk traits of Holstein cows. *J. Dairy Sci.*, 96(3):1521–1534. 2013a.

N. Melzer, D. Wittenburg, and D. Repsilber. 2010b. Simulating snp data: influence of simulation design on the extent of linkage disequilibrium. Pages 19–22. In H. M. Seifert and G. Viereck (eds.), *11th day of the Doctoral Student*. BUK! Breitschuh & Kock GmbH, Kiel, Germany.

N. Melzer, D. Wittenburg, and D. Repsilber. 2011. Including metabolomic profiles to improve genetic value prediction: an integrated bioinformatics approach using weighted genome-wide marker information. Pages 55–58. In H. M. Seifert and G. Viereck (eds.), *12th day of the Doctoral Student*. BUK! Breitschuh & Kock GmbH, Kiel, Germany.

N. Melzer, D. Wittenburg, and D. Repsilber. Simulating SNP data: influence of simulation design on the extent of linkage disequilibrium. *Arch. Tierz.*, 56(38):380–398. 2013b.

N. Melzer, D. Wittenburg, and D. Repsilber. Analyzing milk metabolite profiles to enable prediction of traditional milk traits for Holstein cows based on SNP information. *PLoS ONE*, 8(8):e70256. 2013c.

P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(Suppl 2):ii122–ii129. 2003.

T. Meuwissen. Genomic selection: marker assisted selection on a genome wide scale. *J. Anim. Breed. Genet.*, 124(6):321–322. 2007.

T. Meuwissen, B. Hayes, and M. Goddard. Accelerating improvement of livestock with genomic selection. *Annu. Rev. Anim. Biosci.*, 1(1):221–237. 2013.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829. 2001.

T. H. E. Meuwissen, T. R. Solberg, R. Shepherd, and J. A. Woolliams. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.*, 41:2. 2009.

T. H. E. Meuwissen and J. A. M. Van Arendonk. Potential improvements in rate of genetic gain from marker-assisted selection in dairy cattle breeding schemes. *J. Dairy Sci.*, 75(6):1651–1659. 1992.

K. Meyn. Entwicklung, Stand und Perspektiven der Rinder- und Schweineproduktion. *Züchtungskunde*, 77(6):478–489. 2005.

F. Miglior, A. Sewalem, J. Jamrozik, J. Bohmanova, D. M. Lefebvre, and R. K. Moore. Genetic analysis of milk urea nitrogen and lactose and their relationships with other production traits in canadian holstein cattle. *J. Dairy Sci.*, 90(5):2468–2479. 2007.

J. H. Moore. A global view of epistasis. *Nat. Genet.*, 37(1):13–4. 2005.

N. E. Morton. Linkage disequilibrium maps and association mapping. *J. Clin. Invest.*, 115(6):1425–1430. 2005.

G. Moser, M. S. Khatkar, B. J. Hayes, and H. W. Raadsma. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genet. Sel. Evol.*, 42:37. 2010.

M. O. Mosig, E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller, and A. Friedmann. A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics*, 157(4):1683–1698. 2001.

T. Motyl, T. Płoszaj, A. Wojtasik, W. Kukulska, and M. Podgurniak. Polyamines in cow's and sow's milk. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.*, 111(3):427–433. 1995.

R. Mrode. *Linear Models for the Prediction of Animal Breeding Values*, CABI Publishing, Oxfordshire, UK. 2 ed. 2005.

J. Nadesalingam, Y. Plante, and J. P. Gibson. Detection of QTL for milk production on chromosomes 1 and 6 of Holstein cattle. *Mamm. Genome*, 12(1):27–31. 2001.

T. Nomura. Toward integration of biological and physiological functions at multiple levels. *Front. Physiol.*, 1:164. 2010.

U. Ober, M. Erbe, N. Long, E. Porcu, M. Schlather, and H. Simianer. Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics*, 188(3):695–708. 2011.

P. A. Oltenacu and D. M. Broom. The impact of genetic selection for increased milk yield on the welfare of dairy cows. *Anim. Welfare*, 19(S):39–49. 2010.

J. D. Orth, I. Thiele, and B. Ø. Palsson. What is flux balance analysis? *Nat. Biotechnol.*, 28(3):245–248. 2010.

B. Palsson. The challenges of in silico biology. *Nat. Biotechnol.*, 18(11):1147–1150. 2000.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and the R Core team, nlme: Linear and Nonlinear Mixed Effects Models. 2009. R package version 3.1-96.

A. Pinna, N. Soranzo, I. Hoeschele, and A. de la Fuente. Simulating systems genetics data with SysGenSIM. *Bioinformatics*, 27(17):2459–2462. 2011.

Y. Plante, J. P. Gibson, J. Nadesalingam, H. Mehrabani-Yeganeh, S. Lefebvre, G. Vandervoort, and G. B. Jansen. Detection of quantitative trait loci affecting milk production traits on 10 chromosomes in Holstein cattle. *J. Dairy Sci.*, 84(6):1516–1524. 2001.

K. S. Pollard, H. N. Gilbert, Y. Ge, S. Taylor, and S. Dudoit, multtest: Resampling-based multiple hypothesis testing. 2010. R package version 2.4.0.

A. Polynikis, S. J. Hogan, and M. di Bernardo. Comparing different ODE modelling approaches for gene regulatory networks. *J. Theor. Biol.*, 261(4):511–530. 2009.

E. Ptak and L. R. Schaeffer. Use of test day yields for genetic evaluation of dairy sires and cows. *Livest. Prod. Sci.*, 34(1-2):23–34. 1993.

S. Qanbari, E. C. G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. R. Sharifi, and H. Simianer. The pattern of linkage disequilibrium in german holstein cattle. *Anim. Genet.*, 41(4):346–356. 2010.

R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2008. URL `http://www.R-project.org`, ISBN 3-900051-07-0, R version 2.7.0.

R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2010. URL `http://www.R-project.org`, ISBN 3-900051-07-0, R version 2.11.0.

R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2011. URL `http://www.R-project.org`, ISBN 3-900051-07-0, R version 2.13.2.

W. M. Rauw, E. Kanis, E. N. Noordhuizen-Stassen, and F. J. Grommers. Undesirable side effects of selection for high production efficiency in farm animals: a review. *Livest. Prod. Sci.*, 56(1):15–33. 1998.

R. Reents and F. Reinhardt. Molekulargenetische Informationen als Ergänzung zur Populationsgenetik. *Züchtungskunde*, 79(1):41–45. 2007.

F. Reinhardt, Z. Liu, S. Seefried, S. Rensing, and R. Reents. Genomische Selektion: Stand der Implementierung bei Deutschen Holsteins. *Züchtungskunde*, 83(4/5):248–256. 2011.

S. L. Rodriguez-Zas, B. R. Southey, D. W. Heyen, and H. A. Lewin. Detection of quantitative trait loci influencing dairy traits using a model for longitudinal data. *J. Dairy Sci.*, 85(10):2681–2691. 2002.

U. Roessner and J. Bowne. What is metabolomics all about? *BioTechniques*, 46(5):363–365. 2009.

M. Ron, E. Feldmesser, M. Golik, I. Tager-Cohen, D. Kliger, V. Reiss, R. Domochovsky, O. Alus, E. Seroussi, E. Ezra, and J. I. Weller. A complete genome scan of the Israeli Holstein population for quantitative trait loci by a daughter design. *J. Dairy Sci.*, 87(2):476–490. 2004.

M. Ron and J. I. Weller. From QTL to QTN identification in livestock–winning by points rather than knock-out: a review. *Anim. Genet.*, 38(5):429–439. 2007.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65. 1987.

E. Ruppin, J. A. Papin, L. F. de Figueiredo, and S. Schuster. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr. Opin. Biotechnol.*, 21(4):502–510. 2010.

L. Sachs. *Angewandte Statistik: Anwendung statistischer Methoden*, Springer, Berlin, Germany. 11 ed. 2004.

K. Saito and F. Matsuda. Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol.*, 61:463–489. 2010.

N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, H.-E. Wichmann, J. H. Barrett, I. R. König, S. E. Stevens, S. Szymczak, D.-A. Tregouet, M. M. Iles, F. Pahlke, H. Pollard, W. Lieb, F. Cambien, M. Fischer, W. Ouwehand, S. Blankenberg, A. J. Balmforth, A. Baessler, S. G. Ball, T. M. Strom, I. Brænne, C. Gieger, P. Deloukas, M. D. Tobin, A. Ziegler, J. R. Thompson, and H. Schunkert. Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.*, 357(5):443–453. 2007.

J. Sanguansermsri, P. György, and F. Zilliken. Polyamines in human and cow's milk. *Am. J. Clin. Nutr.*, 27(8):859–65. 1974.

K. Sax. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus Vulgaris*. *Genetics*, 8(6):552–560. 1923.

J. M. Schefers and K. A. Weigl. Genomic selection in dairy cattle: integration of DNA testing into breeding programs. *Animal Frontiers*, 2(1):4–9. 2012.

L. Scheubert, M. Luštrek, R. Schmidt, D. Repsilber, and G. Fuellen. Tissue-based Alzheimer gene expression markers - comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. *BMC Bioinformatics*, 13:266. 2012.

R. D. Schnabel, T. S. Sonstegard, J. F. Taylor, and M. S. Ashwell. Whole-genome scan to detect QTL for milk production, conformation, fertility and functional traits in two US Holstein families. *Anim. Genet.*, 36(5):408–416. 2005.

C. Schrooten, M. C. A. M. Bink, and H. Bovenhuis. Whole genome scan to detect chromosomal regions affecting multiple traits in dairy cattle. *J. Dairy Sci.*, 87(10):3550–3560. 2004.

Y. H. Schukken, D. J. Wilson, F. Welcome, L. Garrison-Tikofsky, and R. N. Gonzalez. Monitoring udder health and milk quality using somatic cell counts. *Vet. Res.*, 34(5):579–596. 2003.

R. Schuster and H.-G. Holzhütter. Use of mathematical models for predicting the metabolic effect of large-scale enzyme activity alterations. application to enzyme deficiencies of red blood cells. *Eur. J. Biochem.*, 229(2):403–418. 1995.

D. F. Schwarz, I. R. König, and A. Ziegler. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, 26(14):1752–1758. 2010. Version 0.7.2.

J. Schwender. *Plant Metabolic Networks*, Springer, New York, USA. 2009.

F. Seefried, Z. Liu, G. Thaller, and F. Reinhardt. Die genomische Zuchtwertschätzung bei der Rasse Deutsche Holstein. *Züchtungskunde*, 82(1):14–21. 2010.

S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, 82(398):605–610. 1987.

M. Slatkin. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, 9(6):477–485. 2008.

D. Sorensen and D. Gianola. *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*, Springer, New York, USA. Statistics for Biology and Health. 2002.

R. J. Spelman, W. Coppieters, L. Karim, J. A. M. van Arendonk, and H. Bovenhuis. Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics*, 144(4):1799–1808. 1996.

R. J. Spelman, M. Keehan, V. Obolonkin, and W. Coppieters. 2007. Application of genomic information in a dairy cattle breeding scheme. In *Proceedings of the Association for the Advancement of Animal Breeding and Genetics (AAABG)*, vol. 17.

J. L. Spratlin, N. J. Serkova, and S. G. Eckhardt. Clinical applications of metabolomics in oncology: a review. *Clin.Cancer Res.*, 15(2):431–440. 2009.

W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcaMethods – a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23:1164–1167. 2007.

A. H. Sturtevant. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *J. Exp. Zool.*, 14:43–59. 1913.

M. Sugimoto, M. Kawakami, M. Robert, T. Soga, and M. Tomita. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr. Bioinform.*, 7(1):96–108. 2012.

J. D. Sutton and S. V. Morant. A review of the potential of nutrition to modify milk fat and protein. *Livest. Prod. Sci.*, 23(3-4):219–237. 1989.

M. Terzer, N. D. Maynard, M. W. Covert, and J. Stelling. Genome-scale metabolic networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 1(3):285–297. 2009.

The Bovine Genome Sequencing and Analysis Consortium, C. G. Elsik, R. L. Tellam, and K. C. Worley. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science*, 324(5926):522–528. 2009.

The phpMyAdmin Project. 2006. URL `www.phpmyadmin.net`, Version 2.9.1.1.

Thermo Fisher Scientific, NanoDrop 1000 Spectrophotometer. Thermo Fisher Scientific Int., Wilmington, USA. 2008. V3.7 User's Manual.

A. Töpel. *Chemie und Physik der Milch*, Behr's Verlag, Hamburg, Germany. 3 ed. 2004.

M. A. Toro and L. Varona. A note on mate allocation for dominance handling in genomic selection. *Genet. Sel. Evol.*, 42:33. 2010.

United States Department of Agriculture. 2008. URL `http://www.marc.usda.gov/genome/cattle/cattle.html`, accessed Nov 2008.

S. van der Beek. Effect of genomic selection on national and international genetic evaluations. *Interbull Bull.*, 37:115–118. 2007.

P. M. VanRaden, C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. Invited review: reliability of genomic predictions for North American Holstein bulls. 92(1):16–24. 2009.

A. I. Vazquez, G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. *J. Dairy Sci.*, 93(12):5942 – 5949. 2010.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*, 4 ed., Springer, New York, USA. Statistics and Computing. 2002.

L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka. 2009. On the comparison of relative clustering validity criteria. Pages 733–744. In *Proceedings of the 2009 SIAM International Conference on Data Mining*. Sparks, Nevada, USA.

K. L. Verbyla, P. J. Bowman, B. J. Hayes, and M. E. Goddard. Sensitivity of genomic selection to using different prior distributions. *BMC Proc.*, 4(Suppl 1):S5. 2010.

K. L. Verbyla, B. J. Hayes, P. J. Bowman, and M. E. Goddard. Accuracy of genomic selection using stochastic search variable selection in australian holstein friesian dairy cattle. *Genet. Res. (Camb.)*, 91(5):307–311. 2009.

C. Viguier, S. Arora, N. Gilmartin, K. Welbeck, and R. O'Kennedy. Mastitis detection: current trends and future perspectives. *Trends Biotechnol.*, 27(8):486–493. 2009.

P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era–concepts and misconceptions. *Nat. Rev. Genet.*, 9(4):255–266. 2008.

B. Vlaeminck, C. Dufour, A. M. van Vuuren, A. R. J. Cabrita, R. J. Dewhurst, D. Demeyer, and V. Fievez. Use of odd and branched-chain fatty acids in rumen contents and milk as a potential microbial marker. *J. Dairy Sci.*, 88(3):1031–1042. 2005.

W. Weckwerth. Metabolomics in systems biology. *Annu. Rev. Plant Biol.*, 54:669–689. 2003.

K. A. Weigel, G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J. Dairy Sci.*, 92(10):5248 – 5257. 2009.

J. Weiß, W. Pabst, and S. Granz. *Tierproduktion*, Enke Verlag, Stuttgart, Germany. 14 ed. 2011.

J. I. Weller and M. Ron. Invited review: Quantitative trait nucleotide determination in the era of genomic selection. *J. Dairy Sci.*, 94(3):1082–1090. 2011.

H. V. Westerhoff and B. O. Palsson. The evolution of molecular biology into systems biology. *Nat. Biotechnol.*, 22(10):1249–1252. 2004.

W. Wiechert. Modeling and simulation: tools for metabolic engineering. *J. Biotechnol.*, 94(1):37–63. 2002.

D. Wittenburg, N. Melzer, and N. Reinsch. 2010. Including non-additive effects in Bayesian methods for the prediction of genetic values from genome-wide SNP data. Abstract ID 0267. In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*. Gesellschaft für Tierzuchtwissenschaften e.V., Leipzig, Germany.

D. Wittenburg, N. Melzer, and N. Reinsch. Including non-additive genetic effects in bayesian methods for the prediction of genetic values based on genome-wide markers. *BMC Genet.*, 12:74. 2011.

D. Wittenburg, N. Melzer, L. Willmitzer, J. Lisec, U. Kesting, N. Reinsch, and D. Repsilber. Milk metabolites and their genetic variability. *J. Dairy Sci.*, 96(4):2557–2569. 2013.

D. Wittenburg and N. Reinsch. 2011. Application of spike and slab variable selection for the genome-wide estimation of genetic effects and their complexity. In *Book of Abstracts of the 62nd Annual Meeting of the EAAP*. Wageningen Academic Publishers, The Netherlands, p. 116. ISSN 1382-6077.

H. Wold. 1975. Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In J. Gani (ed.), *Perspectives in Probability and Statistics, Papers in honour of M. S. Barlett.* Academic Press, London, Pages 117–142.

O. Wolkenhauer. Systems biology: the reincarnation of systems theory applied in biology? *Brief. Bioinform.*, 2(3):258–270. 2001.

O. Wolkenhauer. Defining systems biology: an engineering perspective. *IET Syst. Biol.*, 1(4):204–206. 2007.

O. Wolkenhauer, D. Shibata, and M. D. Mesarović. The role of theorem proving in systems biology. *J. Theor. Biol.*, 300(0):57–61. 2012.

O. Wolkenhauer, M. Ullah, P. Wellstead, and K.-H. Cho. The dynamic systems approach to control and regulation of intracellular networks. *FEBS Lett.*, 579(8):1846–1853. 2005.

S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159. 1931.

J. Xiong. Cambridge University Press, New York. *Essential Bioinformatics*. 2006.

J.-H. Xue and D. M. Titterington. Do unbalanced data have a negative effect on LDA? *Pattern Recogn.*, 41(5):1575–1588. 2008.

S. P. Yadav. The wholeness in suffix -omics, -omes, and the word om. *J. Biomol. Tech.*, 18(5):277. 2007.

Q. Zhang, D. Boichard, I. Hoeschele, C. Ernst, A. Eggen, B. Murkve, M. Pfister-Genskow, L. A. Witte, F. E. Grignola, P. Uimari, G. Thaller, and M. D. Bishop. Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics*, 149(4):1959–1973. 1998.

Z. Zhang, J. Liu, X. Ding, P. Bijma, D.-J. de Koning, and Q. Zhang. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One*, 5(9):e12648. 2010.

A. Ziegler, I. R. König, and J. R. Thompson. Biostatistical aspects of genome-wide association studies. *Biom. J.*, 50(1):8–28. 2008.

O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA*, 109(4):1193–1198. 2012.

Appendix A

Additional information for fastBayesB

A.1 γ-values which were not suitable for fastBayesB

Table A.1: The table items represent the number of replicates leading to non-convergence and aborting rates over 100 replicates for each tested scenario (S) where they occurred for the conventional approach (Conv) and the SBML approach. The following scenarios were tested: (1) $n_{SNP} = 5{,}227$ and $n_{QTL} = 23$; (2) $n_{SNP} = 5{,}227$ and $n_{QTL} = 230$; (3) $n_{SNP} = 52{,}273$ and $n_{QTL} = 23$; (4) $n_{SNP} = 52{,}273$ and $n_{QTL} = 230$. In addition zero means that fastBayesB converged for all 100 replicates.

			Not converged			Abort		
Approach	S	γ	$H^2 = 0.1$	$H^2 = 0.3$	$H^2 = 0.5$	$H^2 = 0.1$	$H^2 = 0.3$	$H^2 = 0.5$
Conv	1	0.1	1	0	34	0	0	0
Conv	1	1e-05	0	0	0	9	0	0
Conv	2	0.1	0	6	24	0	0	0
Conv	2	1e-05	0	0	0	18	0	0
Conv	3	0.1	0	6	4	100	79	79
Conv	3	0.05	100	1	0	0	26	65
Conv	3	0.025	41	0	0	0	0	0
Conv	3	0.01	1	0	0	0	0	0
Conv	4	0.1	0	2	2	100	89	82
Conv	4	0.05	100	0	0	0	34	74
Conv	4	0.025	50	0	0	0	0	0
Conv	4	0.005	0	3	0	0	0	0
SBML	1	0.1	0	1	32	0	0	0
SBML	1	1e-05	0	0	0	13	1	0
SBML	2	0.1	0	3	26	0	0	0
SBML	2	1e-05	0	0	0	25	0	0
SBML	3	0.1	0	4	9	99	82	74
SBML	3	0.05	100	0	0	0	16	49
SBML	3	0.025	38	0	0	0	0	0
SBML	3	0.01	1	0	0	0	0	0
SBML	3	0.005	0	0	1	0	0	0
SBML	4	0.1	0	3	2	100	82	85
SBML	4	0.05	100	0	1	0	34	77
SBML	4	0.025	60	0	0	0	0	0

A.2 Goodness of model fit for simulated data

Table A.2: For the optimal γ-value, the mean correlation between fitted values and residuals is listed for all tested scenarios for the conventional approach and SBML approach. In parentheses the corresponding standard deviations are presented for 100 replicates. The following scenarios were tested: (1) n_{SNP} = 5,227 and $n_{QTL} = 23$; (2) n_{SNP} = 5,227 and $n_{QTL} = 230$; (3) n_{SNP} = 52,273 and $n_{QTL} = 23$; (4) n_{SNP} = 52,273 and $n_{QTL} = 230$.

Scenario	Conventional approach	SBML approach
	$H^2 = 0.1$	
1	0.054 (0.013)	0.056 (0.010)
2	0.520 (0.051)	0.557 (0.039)
3	0.069 (0.018)	0.069 (0.016)
4	0.693 (0.049)	0.730 (0.030)
	$H^2 = 0.3$	
1	0.022 (0.005)	0.025 (0.005)
2	0.087 (0.022)	0.137 (0.024)
3	0.028 (0.007)	0.031 (0.006)
4	0.060 (0.010)	0.198 (0.015)
	$H^2 = 0.5$	
1	0.013 (0.003)	0.047 (0.005)
2	0.073 (0.019)	0.051 (0.021)
3	0.017 (0.004)	0.019 (0.004)
4	0.115 (0.008)	0.129 (0.010)

In general a correlation of zero means that the genetic variance and the residual variance is perfectly separated, i.e., the model assumptions match the truth sufficiently (genetic effects and residuals were assumed independent). As an example, the model explains the data in case of scenario 1 and H^2=0.3 well, and in case of scenario 2 and H^2=0.1 not sufficiently for both simulation approaches.

Appendix B

Additional information about experimental data analyses

B.1 Correlation between milk metabolites and milk traits

The following table shows the Pearson's correlation coefficients between milk metabolites and milk traits and among milk traits themselves. The columns show the milk traits. The first rows of Table B.1 show the milk traits as well, while the latter rows show the milk metabolites.

Table B.1: Pearson correlation matrix of milk metabolites and milk traits.

Trait	Acetone (%)	Casein (%)	Fat (%)	Lactose (%)	pH value	Protein (%)	Quantity of milk (kg)	SFA	SCS	UFA	Urea (%)	Energy (MJ/kg)	Fat/ lactose	Fat/ protein
Milk traits														
Acetone (%)	1.00	0.05	0.28	-0.07	0.02	0.05	-0.01	0.37	0.10	-0.18	-0.15	0.26	0.29	0.27
Casein (%)	0.05	1.00	0.27	0.11	0.11	0.99	-0.34	-0.04	0.06	0.01	-0.11	0.48	0.24	-0.22
Fat (%)	0.28	0.27	1.00	-0.07	0.19	0.26	-0.21	0.78	0.13	0.39	0.02	0.97	0.98	0.88
Lactose (%)	-0.07	0.11	-0.07	1.00	0.50	0.00	0.04	0.00	-0.28	-0.06	0.06	0.04	-0.26	-0.08
pH value (%)	0.02	0.11	0.19	0.50	1.00	0.08	-0.11	0.12	-0.17	-0.24	0.07	0.24	0.09	0.15
Protein (%)	0.05	0.99	0.26	0.00	0.08	1.00	-0.36	-0.06	0.10	0.02	-0.08	0.47	0.26	-0.22
Quantity of milk (kg)	-0.10	-0.34	-0.21	0.04	-0.11	-0.36	1.00	-0.06	-0.06	-0.02	0.11	-0.27	-0.21	-0.03
SFA	0.37	-0.04	0.78	0.00	0.12	-0.06	-0.06	1.00	0.11	0.29	-0.06	0.70	0.76	0.83
SCS	0.10	0.06	0.13	-0.28	-0.17	0.10	-0.06	0.11	1.00	0.08	-0.12	0.11	0.18	0.08
UFA	-0.18	0.01	0.39	-0.06	-0.24	0.02	-0.02	0.29	0.08	1.00	-0.03	0.36	0.40	0.38
Urea (%)	-0.15	-0.11	0.02	0.06	0.07	-0.08	0.11	-0.06	-0.12	-0.03	1.00	0.00	0.01	0.06
Energy	0.26	0.48	0.97	0.04	0.24	0.47	-0.27	0.70	0.11	0.36	0.00	1.00	0.93	0.75
Fat/lactose	0.29	0.24	0.98	-0.26	0.09	0.26	-0.21	0.76	0.18	0.40	0.01	0.93	1.00	0.86
Fat/protein	0.27	-0.22	0.88	-0.08	0.15	-0.22	-0.03	0.83	0.08	0.38	0.06	0.75	0.86	1.00
Milk metabolites														
1,3-Dihydroxyaceton	0.05	0.15	0.19	-0.21	-0.05	0.18	-0.11	0.07	0.37	0.17	-0.09	0.20	0.23	0.11
2-Piperidinecarboxylic acid	0.02	-0.22	-0.05	-0.06	-0.11	-0.21	0.16	0.07	0.02	0.05	-0.01	-0.10	-0.04	0.06
4-Hydroxyphenyl-beta-glucopyranoside	-0.01	0.00	-0.03	0.12	0.01	-0.02	0.04	0.01	-0.03	0.01	-0.03	-0.02	-0.06	-0.02
4-(Methylamino)benzoic acid	0.05	0.08	0.05	-0.21	-0.11	0.12	-0.09	0.01	0.40	0.10	-0.04	0.05	0.09	-0.01
Acetic acid, 4-hydroxyphenyl-	-0.02	0.05	-0.02	-0.03	0.02	0.05	0.00	-0.03	0.06	-0.01	-0.05	-0.01	-0.01	-0.05
Acetonitrile, 4-hydroxyphenyl-	-0.02	0.03	0.02	-0.01	0.02	0.04	-0.02	0.01	-0.01	-0.01	0.01	0.02	0.02	0.00
Aconitic acid, cis-	0.07	-0.01	0.00	0.00	-0.02	-0.02	0.02	0.05	0.03	-0.01	-0.07	-0.01	0.00	0.01
Adenine	0.03	0.04	0.02	-0.02	-0.01	0.04	-0.03	0.02	0.02	0.02	-0.04	0.03	0.02	0.00
Adenosine, 2'-deoxy-	0.08	0.08	0.06	0.05	0.08	0.07	-0.05	0.06	0.02	-0.04	-0.04	0.08	0.05	0.02
Adenosine-5-monophosphate	0.01	-0.03	-0.07	0.01	-0.05	-0.03	0.02	-0.06	0.02	-0.01	0.04	-0.07	-0.07	-0.06
Adenosine, alpha-	-0.04	0.00	0.00	0.00	-0.03	0.00	0.03	-0.01	-0.01	0.03	-0.01	0.00	0.00	0.00

Trait	Acetone (%)	Casein (%)	Fat (%)	Lactose (%)	pH value	Protein (%)	Quantity of milk (kg)	SFA	SCS	UFA	Urea (%)	Energy (MJ/kg)	Fat/ lactose	Fat/ protein
Adipamide	0.05	0.07	0.01	-0.02	-0.03	0.07	-0.02	-0.01	0.14	-0.07	0.02	0.02	0.01	-0.03
Adipic acid, 2-amino-	0.04	-0.28	-0.08	-0.01	-0.05	-0.28	0.13	0.07	-0.02	0.00	0.07	-0.14	-0.07	0.06
Ajmaline	-0.02	0.07	0.04	-0.03	-0.01	0.07	0.03	0.04	0.01	-0.01	-0.02	0.05	0.04	0.01
Alanine	0.02	-0.14	-0.03	-0.10	-0.08	-0.13	0.05	0.04	0.02	0.02	0.01	-0.06	-0.01	0.04
Alanine, beta-	-0.03	0.00	-0.05	-0.14	-0.18	0.02	0.08	-0.01	0.14	0.04	-0.03	-0.06	-0.02	-0.06
Allose	-0.05	0.01	-0.05	-0.02	-0.02	0.02	0.04	-0.05	-0.04	-0.02	0.04	-0.04	-0.05	-0.06
Androst-4-en-3,17-dione, 11beta-hydroxy-	0.03	-0.01	0.00	0.00	0.00	-0.01	0.05	0.06	0.03	-0.07	-0.02	0.00	0.00	0.01
Androst-4-en-3,17-dione, 19-hydroxy-	0.03	0.04	0.02	0.01	-0.01	0.04	-0.04	0.02	0.02	-0.04	0.00	0.03	0.02	0.00
Androst-5-en-17-one, 3beta-hydroxy-	0.02	0.10	0.01	0.02	0.04	0.10	-0.04	0.00	0.01	-0.01	-0.04	0.03	0.00	-0.04
Anthranilic acid	-0.02	0.07	-0.01	0.06	0.01	0.06	0.04	-0.01	-0.02	-0.07	0.04	0.01	-0.02	-0.04
Aphidicolin	-0.02	0.06	0.00	0.00	-0.02	0.06	-0.03	0.01	0.00	0.01	-0.02	0.01	0.00	-0.03
Arabitol	0.01	0.33	0.19	-0.03	0.11	0.34	-0.18	-0.06	0.04	0.03	-0.01	0.25	0.19	0.02
Arginine [-NH3]	0.08	-0.11	-0.02	-0.18	-0.14	-0.08	0.00	0.07	0.22	0.04	-0.04	-0.05	0.02	0.02
Asparagine	0.08	-0.24	-0.06	-0.11	-0.10	-0.23	0.12	0.10	0.18	-0.01	-0.02	-0.11	-0.03	0.06
Aspartic acid	-0.01	-0.22	-0.13	0.01	-0.08	-0.22	0.12	-0.01	0.01	0.01	0.07	-0.17	-0.12	-0.02
Aziridine, N-dansyl-	0.03	0.03	-0.01	0.05	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.00	-0.02	-0.02
Benzoic acid,	0.01	0.11	0.01	-0.07	0.01	0.12	-0.06	-0.01	0.08	0.03	0.01	0.03	0.03	-0.04
Benzoic acid, 4-hydroxy-	-0.02	0.00	0.00	-0.03	0.03	0.00	-0.01	-0.01	0.01	0.01	-0.05	0.00	0.01	0.00
Benzylamine	0.02	-0.05	0.05	0.00	0.00	-0.05	0.01	0.11	0.02	0.01	-0.02	0.04	0.05	0.08
beta-Alanyl-histidine	-0.02	0.09	-0.03	0.02	-0.03	0.08	-0.04	-0.03	0.04	-0.04	-0.01	0.00	-0.03	-0.06
beta-D-Fructofuranosyl-(2,1)-beta-D-Fructofuranose	0.00	-0.02	-0.04	0.01	-0.05	-0.03	0.05	-0.01	0.08	0.01	-0.04	-0.04	-0.04	-0.02
beta-D-Galactopyranoside, 1-isopropyl-, 1-thio-	-0.02	0.10	-0.03	0.01	0.04	0.10	-0.02	-0.05	0.01	-0.07	-0.05	-0.01	-0.03	-0.09
Butanoic acid, 2-amino-	0.10	-0.23	-0.03	-0.02	-0.04	-0.24	0.18	0.11	0.08	-0.01	-0.08	-0.08	-0.03	0.09
Butanoic acid, 2-hydroxy-	0.06	-0.03	0.02	-0.18	-0.08	0.00	-0.02	0.04	0.41	0.01	-0.05	0.00	0.05	0.03
Butanoic acid, 3-hydroxy-	0.17	-0.14	0.06	-0.10	0.05	-0.14	0.03	0.14	0.13	0.00	-0.02	0.01	0.08	0.14
Butanoic acid, 4-acetamido-	0.09	0.00	0.06	-0.24	-0.14	0.04	-0.04	0.07	0.32	0.06	0.06	0.04	0.11	0.05
Butanoic acid, 4-amino-	-0.02	-0.09	-0.09	-0.04	-0.07	-0.08	0.05	-0.04	0.04	0.02	0.00	-0.10	-0.08	-0.04
Butanoic acid, 4-hydroxy-	0.07	0.00	0.07	-0.18	-0.08	0.03	-0.05	0.06	0.30	0.07	0.00	0.05	0.10	0.06

Trait	Acetone (%)	Casein (%)	Fat (%)	Lactose (%)	pH value	Protein (%)	Quantity of milk (kg)	SFA	SCS	UFA	Urea (%)	Energy (MJ/kg)	Fat/ lactose	Fat/ protein
Butanoic acid, 4-methylthio-2-oxo-	-0.01	0.03	0.01	0.00	0.02	0.03	0.13	0.03	-0.02	-0.02	0.00	0.02	0.01	-0.01
Butyric acid, 2,4-diamino-, DL-	0.01	-0.02	-0.01	-0.13	-0.08	-0.01	0.00	-0.01	0.23	0.07	-0.05	-0.02	0.02	0.00
Butyro-1,4-lactam, 2-amino-	0.01	0.06	0.02	-0.04	0.04	0.07	-0.04	0.00	0.02	0.01	-0.05	0.03	0.03	-0.01
Cholestane, 3beta-hydroxy-, 5alpha-	0.00	-0.02	-0.02	0.06	0.01	-0.03	0.02	0.00	-0.04	0.00	0.02	-0.02	-0.04	-0.01
Cinnamic acid, 2,5-dimethoxy-, trans-	0.00	0.07	0.03	-0.01	0.02	0.07	-0.03	-0.01	0.04	-0.02	0.00	0.04	0.03	-0.01
Cinnamic acid, 3,4,5-trimethoxy-, trans-	0.01	0.29	-0.01	0.04	0.09	0.30	-0.14	-0.10	0.07	-0.10	0.00	0.06	-0.01	-0.15
Cinnamic acid, 4-hydroxy-, trans-	0.01	0.13	-0.01	0.00	0.01	0.13	-0.04	-0.02	0.03	-0.05	-0.05	0.02	-0.01	-0.07
Citric acid, 2-methyl-	-0.08	0.15	-0.02	0.04	0.05	0.15	-0.01	-0.12	-0.02	-0.03	0.03	0.02	-0.03	-0.10
Citrulline	0.02	0.08	0.04	-0.06	-0.06	0.09	0.01	-0.02	0.01	0.09	-0.01	0.05	0.05	-0.01
Coniferylalcohol, trans-	-0.01	0.06	-0.03	-0.03	-0.05	0.07	-0.03	-0.03	0.05	-0.02	0.04	-0.01	-0.02	-0.06
Coniferylaldehyde, trans-	0.06	0.10	-0.01	-0.03	0.01	0.11	-0.09	-0.06	0.10	-0.06	0.02	0.01	0.00	-0.06
Creatinine	0.02	-0.04	0.02	-0.02	0.02	-0.04	0.05	0.06	0.07	0.01	-0.05	0.01	0.02	0.04
Cystine	0.03	0.06	0.07	0.01	0.00	0.06	-0.04	0.07	0.02	0.00	-0.02	0.08	0.07	0.04
Cytosine	0.02	-0.04	-0.03	-0.02	-0.04	-0.03	0.03	0.02	-0.02	0.04	0.02	-0.03	-0.02	-0.01
Dihydrosphingosine	-0.03	-0.04	-0.03	-0.05	-0.02	-0.04	0.03	-0.01	-0.01	0.01	-0.03	-0.04	-0.02	-0.01
Estradiol, 17alpha-	0.05	-0.09	-0.01	0.04	0.03	-0.09	0.02	0.06	-0.02	0.01	0.01	-0.03	-0.02	0.04
Ethanesulfonic acid, 2-cyclohexylamino-	0.01	0.02	0.04	0.00	0.02	0.02	-0.04	0.04	0.04	0.03	-0.03	0.04	0.04	0.03
Ethanolaminephosphate	0.21	0.03	-0.06	0.00	-0.05	0.03	0.06	-0.02	0.06	-0.07	-0.04	-0.05	-0.06	-0.07
Ethylenediaminetetraacetic acid	-0.01	-0.03	0.01	-0.08	0.00	-0.02	0.00	0.02	0.02	-0.02	-0.04	0.00	0.03	0.02
Etiocholan-17beta-ol-3-one	0.00	0.00	0.03	0.04	0.01	0.00	-0.02	0.04	0.02	0.05	0.01	0.03	0.02	0.03
Fructose-1-phosphate	0.01	0.08	0.04	-0.07	-0.01	0.09	-0.05	0.01	0.09	-0.02	-0.03	0.05	0.06	0.00
Fructose-6-phosphate	0.00	0.16	0.03	0.03	0.06	0.17	-0.06	-0.01	-0.02	-0.05	-0.05	0.07	0.02	-0.06
Fumaric acid	0.07	-0.01	0.09	-0.05	0.01	-0.02	0.00	-0.01	0.00	-0.02	0.03	0.07	0.09	0.09
Fumaric acid, 2-methyl-	0.03	-0.03	0.01	-0.02	0.01	-0.03	-0.02	0.03	0.01	0.02	-0.05	0.00	0.01	0.02
Galactinol	0.00	0.03	-0.03	0.04	0.04	0.03	-0.02	-0.02	0.00	-0.01	-0.07	-0.02	-0.04	-0.04
Galactitol	-0.06	0.03	-0.14	0.08	-0.03	0.02	0.04	-0.11	-0.09	-0.03	-0.10	-0.11	-0.15	-0.15
Galactopyranoside, 1-O-methyl-, beta-	0.07	0.03	0.00	-0.01	-0.03	0.02	-0.02	0.04	0.11	0.02	-0.08	0.01	0.01	0.00
Galactosamine, N-acetyl-	0.02	-0.03	-0.02	-0.05	-0.02	-0.04	0.01	-0.02	0.12	0.03	-0.02	-0.04	-0.01	0.00
Galacturonic acid	0.05	0.12	0.04	-0.06	-0.01	0.13	-0.07	-0.02	0.06	0.02	-0.02	0.06	0.05	-0.02

Trait	Acetone (%)	Casein (%)	Fat (%)	Lactose (%)	pH value	Protein (%)	Quantity of milk (kg)	SFA	SCS	UFA	Urea (%)	Energy (MJ/kg)	Fat/ lactose	Fat/ protein
Galacturonic acid-1-phosphate	-0.03	-0.03	-0.02	-0.06	0.02	-0.01	0.02	-0.02	-0.03	-0.03	-0.03	-0.02	0.00	-0.01
Glucaric acid-1,4-lactone	-0.08	0.03	-0.12	0.14	-0.05	0.01	0.05	-0.07	-0.12	-0.01	-0.08	-0.10	-0.15	-0.13
Gluconic acid	0.05	0.16	0.07	-0.05	0.02	0.17	-0.10	-0.03	0.06	0.03	-0.02	0.10	0.08	-0.01
Gluconic acid-6-phosphate	0.01	0.13	0.06	0.02	0.05	0.14	-0.06	0.03	-0.03	-0.08	-0.04	0.08	0.05	-0.01
Glucopyranoside, 1-O-methyl-, alpha-	0.01	0.07	0.01	0.01	0.03	0.07	0.03	0.02	-0.03	-0.04	-0.10	0.02	0.00	-0.03
Glucosamine, N-acetyl-	0.01	0.05	-0.06	-0.05	-0.10	0.05	-0.01	-0.03	0.14	0.03	-0.06	-0.05	-0.04	-0.08
Glucose, 1,6-anhydro, beta-	0.02	0.05	0.03	0.00	0.01	0.05	-0.04	0.02	-0.06	-0.02	-0.05	0.04	0.03	0.00
Glucose, 2-amino-2-deoxy-	0.03	0.03	0.01	-0.12	-0.09	0.03	-0.05	0.01	0.15	0.03	-0.06	0.00	0.03	-0.01
Glutaric acid, 2-hydroxy-	0.04	0.14	0.08	0.01	0.11	0.14	-0.03	-0.01	0.02	-0.04	-0.06	0.10	0.07	0.02
Glutaric acid, 2-oxo-	0.04	-0.01	0.02	-0.03	0.08	-0.01	0.02	-0.07	-0.02	-0.03	-0.01	0.01	0.02	0.02
Glyceric acid-2-phosphate	-0.02	0.09	0.00	0.01	-0.01	0.08	0.00	-0.02	-0.03	-0.03	0.01	0.02	0.00	-0.04
Glyceric acid-3-phosphate	-0.05	0.23	0.06	0.06	0.05	0.23	-0.07	-0.04	-0.06	-0.08	0.02	0.12	0.05	-0.06
Glycerol	0.01	0.07	-0.01	-0.08	-0.01	0.09	-0.04	-0.03	0.20	0.11	-0.10	0.01	0.01	-0.05
Glycerol-2-phosphate	-0.01	0.17	-0.05	-0.04	-0.15	0.19	-0.07	-0.07	0.07	0.04	-0.05	-0.01	-0.04	-0.15
Glycerol-3-phosphate	-0.01	0.28	-0.03	0.00	-0.13	0.30	-0.11	-0.08	0.07	0.04	-0.06	0.04	-0.03	-0.18
Glycine	0.07	-0.18	-0.06	-0.15	-0.18	-0.16	0.13	0.11	0.10	0.04	-0.09	-0.11	-0.03	0.02
Guanine	-0.03	-0.02	-0.01	0.03	0.02	-0.02	0.04	0.00	-0.06	-0.04	0.03	-0.01	-0.02	-0.01
Guanosine, 2'-deoxy-	-0.03	0.02	-0.06	0.03	-0.01	0.02	0.03	-0.05	-0.03	-0.03	0.02	-0.05	-0.06	-0.06
Hippuric acid	0.00	0.02	-0.01	0.05	0.05	0.02	0.05	-0.02	-0.03	-0.04	0.06	0.00	-0.02	-0.02
Hippuric acid, 2-hydroxy-	0.03	0.16	0.10	-0.06	0.01	0.16	-0.08	-0.02	0.05	-0.01	0.02	0.12	0.11	0.01
Histidine	0.05	0.07	0.01	-0.17	-0.08	0.10	-0.04	0.05	0.31	0.02	-0.05	0.01	0.04	-0.03
Homoserine lactone, N-2-oxocaproyl-	0.04	0.08	0.00	0.04	0.06	0.08	-0.04	0.00	0.02	-0.02	0.01	0.02	-0.01	-0.04
Homoserine lactone, N-tetradecanoyl-	-0.01	0.07	0.00	0.05	0.00	0.06	-0.02	-0.01	0.01	-0.01	-0.04	0.02	-0.01	-0.03
Hydantoin, 5-propionate-	0.03	0.07	-0.03	0.02	0.00	0.06	-0.05	-0.03	0.05	-0.05	0.00	-0.01	-0.04	-0.07
Hydroquinone	0.01	0.06	0.04	-0.03	0.03	0.07	-0.01	0.03	0.05	0.00	-0.01	0.05	0.05	0.01
Imidazole-4-acetic acid, 1-methyl-	-0.01	0.02	-0.01	0.02	-0.03	0.02	0.01	0.01	-0.02	0.01	-0.03	0.00	-0.01	-0.02
Indole-3-acetonitrile	-0.01	0.03	-0.01	-0.02	-0.02	0.03	-0.01	-0.03	-0.04	-0.03	0.05	0.00	0.00	-0.02
Inosine	0.05	-0.06	0.00	0.05	-0.01	-0.07	0.02	0.03	0.03	-0.02	-0.01	-0.01	-0.01	0.04
Inosine-5'-monophosphate	0.08	0.09	0.05	0.06	0.07	0.08	0.00	0.04	-0.07	-0.05	-0.03	0.07	0.04	0.01
Inositol, allo-	-0.03	-0.02	0.00	-0.01	-0.01	-0.02	-0.01	0.00	-0.01	-0.03	-0.02	-0.01	0.00	0.00
Isobutanoic acid, 3-amino-	0.00	-0.01	-0.05	0.01	0.01	-0.02	0.05	-0.03	0.02	-0.03	-0.07	-0.05	-0.05	-0.05

Trait	Acetone (%)	Casein (%)	Fat (%)	Lactose (%)	pH value	Protein (%)	Quantity of milk (kg)	SFA	SCS	UFA	Urea (%)	Energy (MJ/kg)	Fat/ lactose	Fat/ protein
Isocitric acid	-0.02	-0.01	-0.08	0.07	-0.02	-0.03	0.03	-0.03	-0.01	-0.04	-0.06	-0.07	-0.09	-0.06
Isoleucine	0.13	-0.04	0.03	-0.18	-0.13	-0.01	-0.04	0.10	0.33	0.03	-0.06	0.00	0.06	0.04
Isomaltose	0.00	0.00	0.00	0.09	0.03	-0.01	-0.01	0.02	-0.01	0.01	-0.06	0.01	-0.01	0.01
Itaconic acid	0.03	-0.02	0.01	-0.03	0.01	-0.02	-0.01	0.03	0.02	0.02	-0.07	0.01	0.02	0.02
Jasmonic acid methyl ester, 2-trans-	0.03	0.09	0.02	0.08	0.06	0.08	0.01	0.02	-0.07	0.05	-0.02	0.04	0.00	-0.02
Kestose, 1-	0.01	0.04	0.02	0.02	0.01	0.03	-0.03	0.00	-0.05	0.00	-0.04	0.03	0.01	0.00
Kynurenine	0.00	0.07	0.00	-0.04	0.01	0.09	-0.05	0.00	0.05	-0.06	0.03	0.02	0.01	-0.04
Lactic acid, 3-(4-hydroxyphenyl)-	0.05	-0.02	-0.03	-0.02	0.05	-0.02	0.06	-0.03	0.02	-0.01	0.00	-0.03	-0.03	-0.02
Lactic acid dimer	-0.02	-0.02	-0.05	-0.06	-0.02	-0.01	0.04	-0.01	0.07	-0.02	0.00	-0.05	-0.04	-0.05
Lactic acid, DL-	0.09	0.04	0.04	-0.22	-0.11	0.07	-0.06	0.03	0.58	0.05	-0.12	0.03	0.09	0.01
Lactitol	0.01	0.01	-0.08	0.02	0.04	0.01	-0.01	-0.07	-0.01	-0.03	-0.01	-0.07	-0.08	-0.08
Lactobionic acid	0.02	0.00	-0.04	0.02	0.01	0.01	0.00	-0.03	-0.04	-0.03	0.00	-0.03	-0.04	-0.05
Lactulose	0.02	-0.01	0.02	0.01	0.02	-0.01	0.01	0.03	-0.04	0.02	-0.03	0.01	0.01	0.02
Leucine	0.11	0.01	0.05	-0.20	-0.11	0.05	-0.08	0.06	0.41	0.06	-0.03	0.03	0.09	0.03
Loganin	0.01	0.08	0.03	0.01	0.03	0.08	-0.02	0.04	0.01	-0.01	-0.02	0.05	0.03	0.00
Lyxose	0.00	0.06	0.03	-0.07	-0.03	0.06	-0.02	0.00	0.18	0.05	-0.06	0.04	0.05	0.00
Maleic acid	-0.01	-0.01	0.04	0.04	0.06	-0.02	0.01	0.03	0.04	0.02	0.04	0.04	0.03	0.05
Malic acid, 3-oxalo-	-0.02	0.02	-0.02	-0.01	-0.05	0.02	0.02	0.00	-0.03	0.02	0.00	-0.01	-0.02	-0.03
Maltotriose	-0.05	0.00	-0.05	0.09	0.00	-0.01	0.02	-0.01	-0.03	0.04	-0.01	-0.04	-0.06	-0.04
Mannosamine, N-acetyl-	0.00	0.14	0.01	0.04	0.04	0.14	-0.03	-0.04	0.05	-0.02	-0.02	0.04	0.00	-0.07
Melezitose	0.03	-0.02	-0.03	0.00	-0.01	-0.02	0.02	-0.02	0.06	0.00	-0.05	-0.03	-0.03	-0.02
Methionine	0.10	0.09	0.06	-0.25	-0.10	0.13	-0.13	0.05	0.43	0.04	-0.01	0.06	0.11	0.00
Methionine, N-formyl-	0.01	0.01	0.00	-0.02	-0.04	0.01	0.04	0.01	0.02	0.02	-0.03	0.00	0.00	0.00
Morin	0.02	0.01	-0.01	0.00	-0.03	0.01	-0.02	0.00	-0.02	0.02	-0.01	0.00	0.00	-0.01
Muramic acid, N-acetyl-	0.03	0.16	0.12	-0.05	0.07	0.17	-0.13	-0.05	0.04	0.02	0.00	0.14	0.12	0.03
myo-Inositol-1-phosphate	0.04	0.24	0.13	-0.09	0.01	0.27	-0.13	0.02	0.12	0.04	-0.06	0.17	0.14	0.00
Myricetin	0.01	0.05	0.00	0.02	0.00	0.05	-0.02	0.02	0.02	0.04	0.00	0.01	0.00	-0.02
Neuraminic acid, N-acetyl-	0.03	-0.05	-0.04	0.00	-0.01	-0.05	-0.03	-0.03	-0.03	-0.04	0.00	-0.05	-0.04	-0.02
Nicotinamide	-0.02	0.05	-0.01	-0.03	-0.02	0.05	0.04	0.01	0.05	0.00	-0.09	0.00	0.00	-0.03
N,N'-Diacetylchitobiose	0.01	0.06	-0.01	-0.01	-0.02	0.06	-0.01	0.02	0.05	-0.02	-0.05	0.01	0.00	-0.03
Norvaline, 3-hydroxy-	-0.01	0.05	0.01	-0.01	0.03	0.05	0.00	-0.01	0.03	-0.04	-0.07	0.02	0.01	-0.02

Trait	Acetone (%)	Casein (%)	Fat (%)	Lactose (%)	pH value	Protein (%)	Quantity of milk (kg)	SFA	SCS	UFA	Urea (%)	Energy (MJ/kg)	Fat/ lactose	Fat/ protein
Orotic acid	0.20	0.04	-0.06	0.02	-0.04	0.03	0.06	-0.02	0.03	-0.07	-0.05	-0.05	-0.06	-0.07
Pantothenic acid, D-	0.01	0.14	0.03	0.04	0.08	0.14	-0.04	0.00	0.05	-0.03	-0.03	0.06	0.02	-0.04
Phenylacetic acid	-0.02	0.04	-0.03	0.00	0.01	0.05	0.01	-0.06	0.00	0.01	0.10	-0.02	-0.03	-0.06
Phenylalanine	0.12	0.00	0.06	-0.26	-0.13	0.05	-0.10	0.05	0.43	0.07	0.01	0.04	0.11	0.04
Phenylglycol, 3,4-dihydroxy-	-0.01	-0.01	0.04	-0.03	0.00	-0.01	0.06	0.06	0.06	0.02	-0.03	0.03	0.04	0.05
Phenylpyruvic acid	0.04	0.04	-0.02	0.05	0.02	0.03	0.03	-0.01	0.01	-0.03	-0.02	0.00	-0.02	-0.03
Phosphoenolpyruvic acid	-0.05	0.24	0.06	0.03	0.07	0.25	-0.09	-0.07	-0.03	-0.07	0.00	0.12	0.05	-0.07
Phosphoric acid	-0.02	0.03	-0.03	0.07	0.05	0.02	0.06	-0.03	0.03	-0.02	-0.05	-0.02	-0.04	-0.05
Piceatannol	0.01	0.05	0.04	-0.02	0.01	0.05	-0.08	0.02	0.01	-0.01	-0.03	0.04	0.04	0.01
Pregn-4-ene-11beta-17alpha-diol-3-20-dione	0.02	0.01	0.00	0.03	-0.01	0.00	0.00	0.03	0.04	-0.02	-0.01	0.00	-0.01	0.00
Pregn-5-ene-3,21-diol-20-one	0.01	0.00	-0.01	0.01	0.04	0.01	0.04	0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.02
Pregnenolone, 17alpha-hydroxy-	-0.01	0.01	0.04	0.02	0.02	0.00	0.02	0.05	-0.05	0.04	0.01	0.04	0.04	0.04
Progesterone	-0.01	0.00	-0.01	0.02	0.00	0.00	0.04	0.00	0.00	0.01	0.02	-0.01	-0.02	-0.01
Progesterone, 11alpha-hydroxy-	0.01	-0.01	0.01	0.00	-0.03	-0.01	0.01	0.00	-0.01	0.02	0.01	0.00	0.00	0.01
Prostaglandin A2	0.04	0.04	-0.01	0.01	0.02	0.04	-0.04	-0.02	0.00	-0.02	-0.03	0.01	-0.01	-0.03
Prostaglandin D2	0.00	-0.01	0.00	0.02	0.00	-0.02	0.01	0.01	0.04	0.03	-0.03	0.00	0.00	0.01
Prostaglandin E1, 6-oxo-	0.02	0.02	-0.03	0.04	-0.01	0.01	0.00	0.00	0.00	-0.01	0.01	-0.02	-0.03	-0.04
Purine, 6-benzylamino-	0.02	0.00	0.04	0.04	0.02	0.00	0.02	0.07	-0.01	0.02	0.01	0.04	0.03	0.05
Putrescine, N-acetyl-	0.04	0.07	0.02	-0.04	0.01	0.08	-0.05	0.00	0.07	-0.02	-0.01	0.04	0.03	-0.02
Pyridine, 2,3-dihydroxy-	0.01	0.04	0.01	0.01	0.03	0.03	0.01	0.01	-0.01	0.00	-0.05	0.02	0.01	-0.01
Pyridine, 3-hydroxy-	-0.02	0.02	-0.03	-0.02	0.00	0.02	0.01	-0.01	0.03	-0.02	-0.04	-0.02	-0.02	-0.04
Pyridoxal	0.03	-0.06	0.00	0.00	-0.04	-0.07	0.11	0.02	0.04	0.07	0.05	-0.02	0.00	0.03
Pyridoxamine	0.04	0.07	0.01	-0.04	-0.03	0.08	-0.07	-0.02	0.07	-0.02	-0.04	0.02	0.01	-0.04
Pyroglutamic acid	0.03	-0.18	-0.12	0.04	-0.03	-0.18	0.13	0.00	0.00	-0.01	0.00	-0.15	-0.12	-0.02
Pyruvic acid	0.00	0.14	0.11	-0.13	0.03	0.15	-0.07	-0.04	0.19	0.03	-0.03	0.12	0.13	0.03
Pyruvic acid, 4-hydroxyphenyl-	0.07	0.03	0.06	-0.17	-0.08	0.06	-0.02	0.07	0.23	0.03	-0.01	0.05	0.09	0.04
Ribulose	-0.01	0.07	0.04	-0.09	-0.04	0.09	0.01	-0.05	-0.02	0.05	0.09	0.05	0.05	-0.01
Ribulose-5-phosphate	-0.06	0.11	0.03	0.04	0.05	0.12	-0.04	-0.04	-0.06	-0.04	-0.02	0.06	0.02	-0.03
Saccharin	0.03	-0.03	0.00	-0.09	-0.06	-0.04	0.03	0.02	0.11	0.04	-0.01	-0.02	0.02	0.02
Saccharopine	0.03	-0.01	-0.04	0.00	-0.02	-0.01	0.00	-0.02	-0.02	-0.03	-0.01	-0.04	-0.04	-0.04

Trait	Acetone (%)	Casein (%)	Fat (%)	Lactose (%)	pH value	Protein (%)	Quantity of milk (kg)	SFA	SCS	UFA	Urea (%)	Energy (MJ/kg)	Fat/ lactose	Fat/ protein
Salicylaldehyde-beta-D-glucopyranoside	0.00	-0.02	-0.01	0.02	-0.02	-0.02	0.04	0.02	0.01	-0.01	0.00	-0.01	-0.01	0.01
Sarcosine	0.05	-0.09	0.01	-0.02	-0.02	-0.09	0.03	0.07	0.02	0.00	-0.06	-0.01	0.02	0.06
Secologanin	-0.01	0.05	-0.03	0.04	0.02	0.04	0.01	-0.03	0.02	0.00	-0.01	-0.01	-0.04	-0.05
Sedoheptulose, 2,7-anhydro-, beta-	-0.07	0.06	-0.10	0.12	-0.02	0.05	0.02	-0.07	-0.14	-0.05	-0.04	-0.07	-0.12	-0.13
Senecionine	-0.03	0.02	-0.04	-0.04	-0.05	0.02	0.01	-0.04	0.03	-0.02	0.00	-0.04	-0.03	-0.05
Serine, cyclo-	0.04	-0.12	0.07	-0.08	-0.12	-0.11	0.09	0.19	0.10	0.10	0.01	0.03	0.08	0.13
Sophorose	0.06	0.01	-0.02	0.09	0.05	0.00	0.03	-0.01	-0.05	-0.03	-0.02	-0.01	-0.04	-0.02
Sorbitol, 1,4:3,6-dianhydro-	-0.02	-0.02	-0.01	-0.04	0.00	-0.02	0.01	0.02	-0.03	0.01	-0.03	-0.01	0.00	0.00
Spermidine	0.08	0.26	0.04	-0.11	0.01	0.28	-0.18	-0.03	0.34	-0.05	-0.06	0.09	0.06	-0.09
Spermine [+CO2]	-0.03	0.04	-0.04	0.02	0.01	0.03	0.00	-0.03	0.01	0.01	-0.02	-0.03	-0.04	-0.05
Sphingosine	-0.01	-0.06	-0.02	-0.08	-0.02	-0.05	0.01	-0.02	-0.01	0.04	-0.03	-0.04	-0.01	0.00
Succinic acid	0.04	-0.02	0.01	-0.08	-0.04	-0.01	0.04	-0.01	0.05	0.01	-0.03	0.00	0.03	0.02
Thiazole, 4-methyl-5-hydroxyethyl-	0.00	0.20	0.04	-0.04	-0.01	0.21	-0.12	-0.05	0.12	0.04	-0.02	0.08	0.05	-0.06
Threonic acid	0.05	0.03	0.00	-0.09	0.00	0.06	-0.07	0.01	0.21	0.00	-0.02	0.00	0.02	-0.02
Thymine	0.04	0.04	0.03	-0.15	-0.10	0.07	-0.03	0.00	0.31	0.10	-0.01	0.03	0.07	0.00
Thymine, 5,6-dihydro-	0.04	-0.10	0.01	-0.06	-0.03	-0.09	0.11	0.11	0.14	0.03	0.01	-0.02	0.02	0.05
Trehalose-6-phosphate	0.02	0.04	0.00	0.02	0.00	0.04	-0.06	-0.01	0.02	0.01	-0.02	0.01	0.00	-0.02
Trehalose, beta,beta'-	0.03	0.01	-0.03	0.05	0.01	0.00	-0.03	-0.04	-0.02	-0.04	-0.01	-0.02	-0.04	-0.03
Tropic acid	0.05	0.10	0.09	0.05	0.12	0.09	-0.04	0.04	0.08	-0.04	-0.03	0.11	0.08	0.05
Tryptamine, 1-methyl-	0.04	0.14	-0.01	-0.03	0.04	0.16	-0.11	-0.07	0.07	-0.05	0.06	0.02	-0.01	-0.09
Tryptophan	0.12	0.03	0.02	-0.21	-0.11	0.07	-0.05	0.05	0.38	0.03	0.00	0.02	0.06	-0.01
Tyrosine	0.10	0.07	0.04	-0.27	-0.13	0.11	-0.11	0.04	0.49	0.06	-0.07	0.04	0.09	0.00
unknown_1031200	0.00	0.03	0.02	0.04	0.04	0.03	-0.01	0.00	0.01	0.00	-0.04	0.03	0.02	0.01
unknown_315800	0.02	0.02	0.04	0.02	0.02	0.02	0.08	0.03	0.03	-0.03	-0.06	0.04	0.03	0.02
unknown_576300	0.02	0.06	-0.01	0.03	0.03	0.06	0.05	-0.02	-0.04	0.00	0.04	0.01	-0.01	-0.03
Uracil	0.04	0.06	0.08	-0.18	-0.12	0.09	-0.05	0.05	0.70	0.11	-0.09	0.07	0.11	0.03
Uracil, dihydro-	0.00	-0.01	-0.02	-0.06	-0.05	0.01	-0.01	-0.01	0.16	0.02	-0.05	-0.02	0.00	-0.02
Uridine	0.03	0.01	0.02	-0.02	0.00	0.01	0.00	0.02	0.05	0.02	-0.05	0.02	0.03	0.02
Uridine-5-monophosphate	-0.02	0.03	-0.03	0.08	0.03	0.02	0.01	-0.03	0.00	0.01	-0.01	-0.02	-0.05	-0.04
Valero-1,5-lactam	0.02	0.01	0.00	0.01	0.02	0.01	0.01	0.04	0.06	-0.01	-0.03	0.00	-0.01	-0.01

B.2 Known QTL regions determined by use of the cattleQTL database

In the following table the genome regions, which were used for the realization of the QTL approach, are specified in detail.

Table B.2: Known QTL regions or QTL peaks, which were filtered from the cattleQTL database (Hu et al., 2007). The following criteria were applied: trait name equal to milk fat percentage and milk protein percentage, analysis type equal to QTL, breed equal to Holstein, and chromosome number and both flanking markers or peak markers had to be available. Based on the marker names it was possible to locate the marker in the physical unit base pair (bp) using the annotation Btau4.2 of the bovine genome from the National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nih.gov//genomes/MapView/Bos_taurus/sequence/BUILD.5.2/initial_release/). Also, the known QTN DGAT1 (Grisart et al., 2004) was considered QTL for both milk traits. The SNP marker ARS-BFGL-NGS-4939 was located directly in the DGAT1 region and was used in the analysis.

					Flanked marker left or peak marker			Marker right			Reference
QTL-ID	Chr.	Marker left	Marker peak	Marker right	Start (bp)	End (bp)	New marker label	Start (bp)	End (bp)	New marker label	
Fat (%) - QTL region											
2506, 2507	1	TGLA49		RM095	3,301,552	3,301,666	RH144714	19,594,554	19,594,685	CA095	Nadesalingam et al. (2001)
3530	2	BMS2519		IDVGA37	126,119,271	126,119,390		132,137,676	132,137,885	IDVGA-37	Ron et al. (2004)
2652	3	TGLA263		HUJ246	40,106,746	40,106,863		57,073,218	57,073,470		Ashwell et al. (2004)
2514	6	BM1329		BM143	26,681,421	26,681,567		44,248,348	44,248,461		Nadesalingam et al. (2001)
2724	6	BM1329		TGLA37	26,681,421	26,681,567		52,551,978	52,552,090		Zhang et al. (1998)
9910	6	BMS1242		BMS518	43,638,368	43,638,475		51,401,343	51,401,489		Gao et al. (2009)
3535	7	BM7160		BMS713	3,519	3,697		8,362,905	8,363,057		Ron et al. (2004)
2673	14	ILSTS039		BMS1678	1,198,415	1,198,657	ILSTS39A	9,188,082	9,188,205		Ashwell et al. (2004)
2732	14	ILSTS011		BM302	11,778,542	11,778,810		33,684,486	33,684,630		Zhang et al. (1998)
3172	14	ILSTS039		CSSM066	1,198,415	1,198,657	ILSTS39A	3,940,258	3,940,454		Kühn et al. (2004)
10306	18	BM7109		ILSTS002	38777457	38777615		42,097,397	42,097,533	MB054	Schrooten et al. (2004)
2726	20	TGLA126		TGLA153	23,795,133	23,795,250		34,306,128	34,306,260		Zhang et al. (1998)
2747	20	TGLA304		TGLA153	12,701,617	12,701,702		34,306,128	34,306,260		Arranz et al. (1998)
2571	26	BM188		BM804	34,594,936	34,595,054		46,107,532	46,107,675		Plante et al. (2001)
2735	26	TGLA22		BM4505	5,697,371	5,697,465		34,061,087	34,061,325		Zhang et al. (1998)
2740	27	RM209		BM1857	18,262,877	18,263,010	CA209	39,065,010	39,065,145		Zhang et al. (1998)
Fat (%) - QTL peak											
2440, 2469	3		BL41		30,141,196	30,141,443					Rodriguez-Zas et al. (2002); Heyen et al. (1999)
2442	3		TGLA263		40,106,746	40,106,863					Heyen et al. (1999)

QTL-ID	Chr.	Marker left	Marker peak	Marker right	Marker left or peak marker start (bp)	Marker left or peak marker end (bp)	Marker left or peak marker new marker label	Marker right start (bp)	Marker right end (bp)	Marker right new marker label	Reference
2446	5		BM315		103,840,880	103,840,987					Heyen et al. (1999)
1545	6		BMS470		59,175,906	59,175,973					Chen et al. (2006)
4909	6		BM415		75,304,180	75,304,332					Ashwell et al. (1998)
10282	6		ILSTS090		13,408,991	13,409,137	BOVILS90				Schrooten et al. (2004)
4915	11		BM304		23,891,271	23,891,380					Ashwell et al. (1998)
10305	13		TGLA23		4,349,594	4,349,725					Schrooten et al. (2004)
2454, 2477	14		ILSTS039		1,198,415	1,198,657	ILSTS39A				Heyen et al. (1999); Rodriguez-Zas et al. (2002)
2457	14		CSSM66		3,940,258	3,940,455					Heyen et al. (1999)
2458	14		BM1508		8,274,119	8,274,224					Heyen et al. (1999)
4923	14		BM6425		73,902,471	73,902,642					Ashwell et al. (1998)
4931	17		BM8125		54,265,266	54,265,381					Ashwell et al. (1998)
2481	21		BM8115		2,094,150	2,094,287					Rodriguez-Zas et al. (2002)
2483	21		ILSTS103		34,036,511	34,036,735	BOVILS103				Rodriguez-Zas et al. (2002)
2484	21		TGLA122		57,556,069	57,556,210					Rodriguez-Zas et al. (2002)
2486	22		INRA194		14,270,694	14,270,841					Rodriguez-Zas et al. (2002)
3541	27		BMS2168		4,223,574	4,223,721					Ron et al. (2004)
Protein (%) - QTL region											
2505	1	TGLA49		RM095	3,301,552	3,301,666	RH144714	19,594,554	19,594,685	CA095	Nadesalingam et al. (2001)
2654	3	BL41		ILSTS029	30,141,196	30,141,443		51,415,436	51,415,595	BOVILS29	Ashwell et al. (2004)
2551	5	BM6026		BP1	7,023,104	7,023,270		14,393,913	14,394,233		Plante et al. (2001)
1508	6	BM1329		BM4311	26,681,428	26,681,544		95,853,652	95,853,754		Kučerová et al. (2006)
2511	6	BM143		BM4528	44,248,348	44,248,461		65,937,339	65,937,580		Nadesalingam et al. (2001)
2661	6	AFR227		BM4311	95,770,317	95,770,427		95,853,652	95,853,754		Ashwell et al. (2004)
2725	6	BM143		TGLA37	44,248,348	44,248,461		52,551,978	52,552,090		Zhang et al. (1998)
10394	6	BM1329	BM143	TGLA37	26,681,421	26,681,567		52,551,978	52,552,090		Spelman et al. (1996)
3534, 3536	7	BMS713		TGLA303	8,362,905	8,363,057		23,934,020	23,934,183		Ron et al. (2004)
2558	9	ILSTS013		BMC701	48,265,866	48,265,986		63,071,394	63,071,667		Plante et al. (2001)
2557	10	TGLA272		CSSM39	92,973,178	92,973,280		95,169,738	95,169,927		Plante et al. (2001)
1697	14	BMC1207		BMS1899	34,222,031	34,222,167		51,234,506	51,234,623		Schnabel et al. (2005)

QTL-ID	Chr.	Marker left	Marker peak	Marker right	Marker left or peak marker start (bp)	Marker left or peak marker end (bp)	Marker left or peak marker new marker label	Marker right start (bp)	Marker right end (bp)	Marker right new marker label	Reference
2675	14	BMS1678		ILSTS011	9,188,082	9,188,205		11,778,542	11,778,810		Ashwell et al. (2004)
10308	18	BM7109		ILSTS002	38,777,457	38,777,615		42,097,397	42,097,533	MB054	Schrooten et al. (2004)
10369	18	IDVGA-55		TGLA227	54,087,816	54,087,987		65,352,279	65,352,368		Schrooten et al. (2004)
2685	20	BM4107		BM5004	43,636,535	43,636,702		65,538,724	65,538,866		Ashwell et al. (2004)
2727	20	TGLA126		TGLA153	23,795,133	23,795,250		34,306,128	34,306,260		Zhang et al. (1998)
2748	20	TGLA304		TGLA153	12,701,617	12,701,702		34,306,128	34,306,260		Arranz et al. (1998)
10329	22	CSSM026		CSSM006	3,375,992	3,376,208		12,815,974	12,816,165		Schrooten et al. (2004)
3540	27	BMS689		CSSM036	29,881,182	29,881,335		35,345,775	35,345,954		Ron et al. (2004)
Protein (%)- QTL peak											
2434	2		TGLA377		26,992,639	26,992,735					Heyen et al. (1999)
2515	2		BMS1126		66,606,285	66,606,440					Mosig et al. (2001)
2441, 2470	3		BL41		30,141,196	30,141,443					Heyen et al. (1999); Rodriguez-Zas et al. (2002)
2443	3		TGLA263		40,106,746	40,106,863					Heyen et al. (1999)
2444, 2472, 2516	3		HUJII77		89,149,091	89,149,301					Heyen et al. (1999); Rodriguez-Zas et al. (2002); Mosig et al. (2001)
2517-9	4		TGLA116		59,714,938	59,715,020					Mosig et al. (2001)
1546, 9914	6		BMS2460		93,998,204	93,998,286					Chen et al. (2006); Gao et al. (2009)
2474, 2521-2, 10394	6		BM143		44,248,348	44,248,461					Rodriguez-Zas et al. (2002); Mosig et al. (2001); Spelman et al. (1996)
2475	6		BMC4203		110,293,378	110,293,542					Rodriguez-Zas et al. (2002)
2520	6		BM1329		26,681,421	26,681,567					Mosig et al. (2001)
2523-4, 4908	6		BM415		75,304,180	75,304,332					Mosig et al. (2001); Ashwell et al. (1998)

QTL-ID	Chr.	Marker left	Marker peak	Marker right	Marker left or peak marker start (bp)	Marker left or peak marker end (bp)	Marker left or peak marker new marker label	Marker right start (bp)	Marker right end (bp)	Marker right new marker label	Reference
2525-9	6		CSN3		88,617,507	88,617,696					Mosig et al. (2001)
2530	6		BP7		96,818,703	96,819,009					Mosig et al. (2001)
9913	6		BMS2508		34,554,721	34,554,825					Gao et al. (2009)
10283	6		ILSTS090		13,408,991	13,409,137	BOVILS90				Schrooten et al. (2004)
2476	7		BMS2258		61,673,743	61,673,872					Rodriguez-Zas et al. (2002)
2531-2	7		ILSTS006		95,524,014	95,524,305	MB057				Mosig et al. (2001)
2533	8		BM711		92,683,013	92,683,189					Mosig et al. (2001)
2534	9		BM4208		88,702,847	88,703,007					Mosig et al. (2001)
10301, 10303	13		TGLA23		4,349,594	4,349,725					Schrooten et al. (2004)
2455	14		ILSTS039		1,198,415	1,198,657	ILSTS39A				Heyen et al. (1999)
2480, 2535-6, 4922	14		BM6425		73,902,471	73,902,642					Rodriguez-Zas et al. (2002); Mosig et al. (2001); Ashwell et al. (1998)
2538	20		UWCA26		71,672,995	71,673,117					Mosig et al. (2001)
2485	21		ILSTS054		59,291,083	59,291,217	BOVILS54				Rodriguez-Zas et al. (2002)
2539-40	21		ETH131		22,766,375	22,766,528	MB005				Mosig et al. (2001)
2487	22		CSSM041		33,533,110	33,533,241					Rodriguez-Zas et al. (2002)
4936	25		BMC3224		35,646,575	35,646,759					Ashwell et al. (1998)
10359	27		CSSM043		30,072,300	30,072,557					Schrooten et al. (2004)
10361	28		BMS362		28,302,023	28,302,152					Schrooten et al. (2004)

B.3 Trace plots for selected SNPs using SVS for three investigated milk traits.

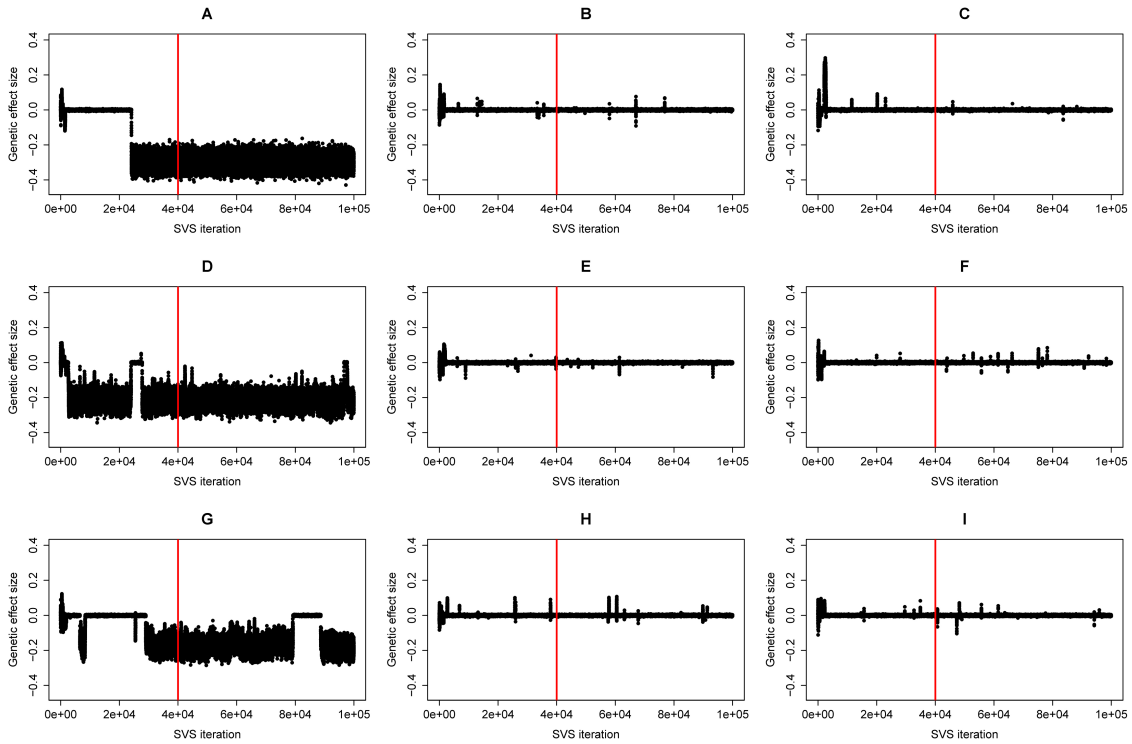


Figure B.1: Trace plots: DGAT1-SNP (A), left SNP position next to DGAT1-SNP (B), right SNP position next to DGAT1-SNP for fat content (C); DGAT1-SNP (D), left SNP position next to DGAT1-SNP (E), right SNP position next to DGAT1-SNP (F) for protein content; DGAT1-SNP (G), left SNP position next to DGAT1-SNP (H), right SNP position next to DGAT1-SNP (I) for pH value. The red line represents the end of the burn-in phase. Figures are based on using the whole data set.

In general a trace plot is a simple graphical tool to study the convergence visually and is commonly used (e.g., Sorensen and Gianola, 2002, pp. 541-550).

B.4 Important milk metabolites detected using RF and PLS

Table B.3: For each milk trait, the observed metabolite importance measurements from a 10-fold cross-validation for the random forest analysis and the partial least squares analysis are listed. The important metabolites fulfilled the condition to have an importance measurement larger than the 90% quantile in each cross-validation run. The order of metabolites implies their importances (descending order), and metabolites typed in bold were detected with both regression methods.

Milk trait	Random forest	Partial least squares
Acetone (%)	**Ethanolaminephosphate**	**Ethanolaminephosphate**
	Orotic acid	**Orotic acid**
	Galactitol	**Glucaric acid-1,4-lactone**
	Glucaric acid-1,4-lactone	Sedoheptulose, 2,7-anhydro-, beta-
		Lyxose
Casein (%)	**Arabitol**	**Arabitol**
	Adipic acid, 2-amino-	**Adipic acid, 2-amino-**
	Cinnamic acid, 3,4,5-trimethoxy-, trans-	**Cinnamic acid, 3,4,5-trimethoxy-, trans-**
	Asparagine	**Glycerol-3-phosphate**
	Glycerol-3-phosphate	**Asparagine**
	Aspartic acid	**Butanoic acid, 2-amino-**
	Pyroglutamic acid	**Aspartic acid**
	Butanoic acid, 2-amino-	**Spermidine**
	2-Piperidinecarboxylic acid	**Phosphoenolpyruvic acid**
	myo-Inositol-1-phosphate	**2-Piperidinecarboxylic acid**
	Alanine	Glyceric acid-3-phosphate
	Spermidine	**myo-Inositol-1-phosphate**
	Thiazole, 4-methyl-5-hydroxyethyl-	**Pyroglutamic acid**
	Phosphoenolpyruvic acid	**Alanine**
		Glycine
		Thiazole, 4-methyl-5-hydroxyethyl-
Fat (%)	**Arabitol**	**Arabitol**
	Galactitol	**1,3-Dihydroxyaceton**
	Pyroglutamic acid	**Galactitol**
	1,3-Dihydroxyaceton	**Aspartic acid**
	Aspartic acid	**Pyroglutamic acid**
	Glucaric acid-1,4-lactone	**myo-Inositol-1-phosphate**
	myo-Inositol-1-phosphate	**Glucaric acid-1,4-lactone**
	Isobutanoic acid, 3-amino-	Muramic acid, N-acetyl-
	Butanoic acid, 4-amino-	Pyruvic acid
	Mannosamine, N-acetyl-	Sedoheptulose, 2,7-anhydro-, beta-
	Fumaric acid	
Lactose (%)	**Tyrosine**	**Tyrosine**
	Phenylalanine	**Phenylalanine**
	Methionine	**Methionine**
	Leucine	Butanoic acid, 4-acetamido-
	1,3-Dihydroxyaceton	Lactic acid, DL-
	Glycine	4-(Methylamino)benzoic acid
	Glucaric acid-1,4-lactone	**Leucine**
		Glucaric acid-1,4-lactone
		Tryptophan
		1,3-Dihydroxyaceton
		Isoleucine

Milk trait	Random forest	Partial least squares
		Arginine [-NH3]
		Butanoic acid, 4-hydroxy-
		Uracil
		Histidine
pH value	**Glycerol-3-phosphate**	**Glycerol-2-phosphate**
	Glycerol-2-phosphate	**Glycerol-3-phosphate**
	Glycine	**Alanine, beta-**
	Arabitol	**Glycine**
	Alanine, beta-	Phenylalanine
	Glucosamine, N-acetyl-	Leucine
	Pantothenic acid, D-	Methionine
		Tyrosine
		Isoleucine
Protein (%)	**Arabitol**	**Arabitol**
	Adipic acid, 2-amino-	**Adipic acid, 2-amino-**
	Glycerol-3-phosphate	**Glycerol-3-phosphate**
	Cinnamic acid, 3,4,5-trimethoxy-, trans-	**Butanoic acid, 2-amino-**
	Asparagine	**Cinnamic acid, 3,4,5-trimethoxy-, trans-**
	myo-Inositol-1-phosphate	**Asparagine**
	Pyroglutamic acid	**Spermidine**
	Butanoic acid, 2-amino-	**Aspartic acid**
	Aspartic acid	**2-Piperidinecarboxylic acid**
	2-Piperidinecarboxylic acid	**Phosphoenolpyruvic acid**
	Spermidine	**myo-Inositol-1-phosphate**
	Thiazole, 4-methyl-5-hydroxyethyl-	**Glyceric acid-3-phosphate**
	Phosphoenolpyruvic acid	**Pyroglutamic acid**
	Glyceric acid-3-phosphate	**Thiazole, 4-methyl-5-hydroxyethyl-**
		Alanine
		Glycerol-2-phosphate
		Glycine
		Butanoic acid, 3-hydroxy-
Quantity of milk (kg)	**Arabitol**	**Butanoic acid, 2-amino-**
	Butanoic acid, 4-methylthio-2-oxo-	Spermidine
	Butanoic acid, 2-amino-	**Arabitol**
	2-Piperidinecarboxylic acid	**2-Piperidinecarboxylic acid**
	Thymine, 5,6-dihydro-	**Butanoic acid, 4-methylthio-2-oxo-**
	Thiazole, 4-methyl-5-hydroxyethyl-	Cinnamic acid, 3,4,5-trimethoxy-, trans-
		Methionine
		Muramic acid, N-acetyl-
		Pyroglutamic acid
		myo-Inositol-1-phosphate
		Glycine
		Pyridoxal
		Aspartic acid
SFA	**Glycerol**	**1,3-Dihydroxyaceto**n
	1,3-Dihydroxyaceton	**Glycerol**
	Spermidine	Uracil
	Glutaric acid, 2-oxo-	4-(Methylamino)benzoic acid
SCS	**Uracil**	**Uracil**
	Lactic acid, DL-	**Lactic acid, DL-**

Milk trait	Random forest	Partial least squares
	Butanoic acid, 2-hydroxy-	**Phenylalanine**
	Tyrosine	**Tyrosine**
	Methionine	**Methionine**
	1,3-Dihydroxyaceton	**Butanoic acid, 2-hydroxy-**
	Cytosine	**Leucine**
	Tryptophan	4-(Methylamino)benzoic acid
	Phenylalanine	**Tryptophan**
	Leucine	**1,3-Dihydroxyaceton**
	Thymine	Butanoic acid, 4-acetamido-
	Sedoheptulose, 2,7-anhydro-, beta-	Isoleucine
		Butanoic acid, 4-hydroxy-
		Histidine
		Spermidine
		Tryptamine, 1-methyl-
		Alanine
UFS	**Serine, cyclo-**	**Serine, cyclo-**
	Galactitol	Butanoic acid, 3-hydroxy-
	Glycerol-3-phosphate	Citric acid, 2-methyl-
	Aconitic acid, cis-	Glycine
		Benzylamine
		Butanoic acid, 2-amino-
		Galactitol
		Asparagine
		Thymine, 5,6-dihydro-
		Isoleucine
		Arabitol
Urea (%)	**Aspartic acid**	Phenylacetic acid
	Glucaric acid-1,4-lactone	**Aspartic acid**
	Tryptophan	Ribulose
	Adipic acid, 2-amino-	Fumaric acid
	Sedoheptulose, 2,7-anhydro-, beta-	**Adipic acid, 2-amino-**
Energy (MJ/kg)	**Arabitol**	**Arabitol**
	Pyroglutamic acid	**1,3-Dihydroxyaceton**
	Aspartic acid	**Aspartic acid**
	myo-Inositol-1-phosphate	**myo-Inositol-1-phosphate**
	1,3-Dihydroxyaceton	**Pyroglutamic acid**
	Galactitol	Adipic acid, 2-amino-
	Thiazole, 4-methyl-5-hydroxyethyl-	Muramic acid, N-acetyl-
	Isobutanoic acid, 3-amino-	Phosphoenolpyruvic acid
	Butanoic acid, 4-amino-	Pyruvic acid
	Glucaric acid-1,4-lactone	Asparagine
		Glyceric acid-3-phosphate
		Hippuric acid, 2-hydroxy-
Fat/protein	**Glycerol-3-phosphate**	**Glycerol-3-phosphate**
	Galactitol	**Butanoic acid, 3-hydroxy-**
	Glycerol-2-phosphate	**Galactitol**
	Butanoic acid, 3-hydroxy-	**Glycerol-2-phosphate**
	1,3-Dihydroxyaceton	Cinnamic acid, 3,4,5-trimethoxy-, trans-
	Glucaric acid-1,4-lactone	Fumaric acid
	Sedoheptulose, 2,7-anhydro-, beta-	Serine, cyclo-
		1,3-Dihydroxyaceton
		Arabitol

Milk trait	Random forest	Partial least squares
		Glucaric acid-1,4-lactone
		Sedoheptulose, 2,7-anhydro-, beta-
		Butanoic acid, 2-amino-
Fat/lactose	**Arabitol**	**1,3-Dihydroxyaceton**
	1,3-Dihydroxyaceton	**Arabitol**
	Galactitol	**Galactitol**
	Pyroglutamic acid	**Glucaric acid-1,4-lactone**
	Glucaric acid-1,4-lactone	Methionine
	myo-Inositol-1-phosphate	Fumaric acid
	Aspartic acid	Phenylalanine
	Isobutanoic acid, 3-amino-	Tyrosine
	Pyruvic acid	**Pyroglutamic acid**
	Thiazole, 4-methyl-5-hydroxyethyl-	4-(Methylamino)benzoic acid
		Pyruvic acid
		Butanoic acid, 4-acetamido-

B.5 Goodness of model fit for experimental data

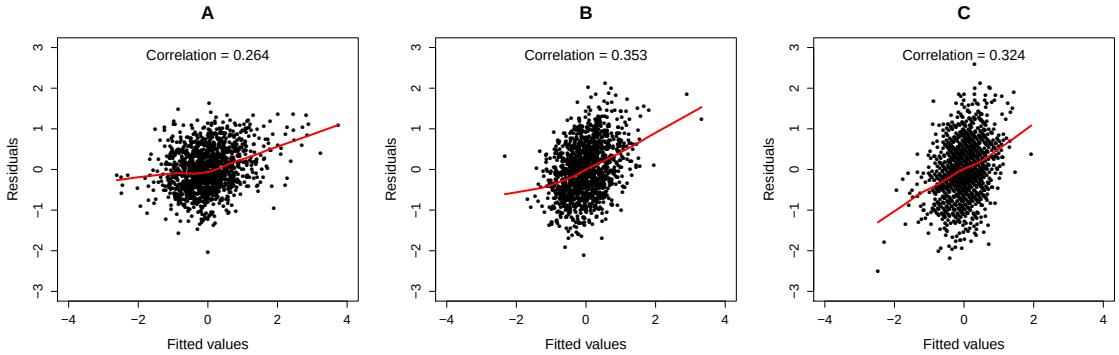


Figure B.2: Residual plots for fat content (A), protein content (B) and pH value (C) using SVS. The obtained correlation between fitted values and residuals is also stated on top of each figure. The red line represents the lowess-smooth of fitted values and residuals. Figures are based on using the whole data set.

In addition, to get an impression of the nature of the relationship between the fitted values and residuals, non-parametric regression can be used. In our case, we used the lowess-smooth (Cleveland, 1979) as implemented in the R (R Development Core Team, 2010) function lowess. Figure B.2 shows that the used linear model considering additive genetic effects does not explain the data completely. It is expected that a linear model, which considers additive and non-additive genetic effect sizes, possibly better fits the data. Another possibility is that the linear model is not adequate, as further unknown influencing factors are contained.

B.6 Important SNP positions detected using the metabolite approach

Table B.4: Important SNP markers, occurring in more than seven cross-validation (CV) runs, for each milk trait obtained via the metabolite approach.

SNP marker	Chromosome	Position in bp	QTL region	QTL peak	Frequency
Fat (%)					
Hapmap42518-BTA-34464	2	11,5611,887			8
Hapmap50895-BTA-122111	4	24,553,482			10
ARS-BFGL-NGS-64882	7	81,392,639			8
ARS-BFGL-NGS-17358	8	102,312,736			9
BTA-63354-no-rs	10	33,150,420			9
Hapmap58072-rs29010006	12	63,149,779			8
BTB-01123944	13	6,281,252			8
ARS-BFGL-BAC-11928	13	29,191,010			9
ARS-BFGL-NGS-57820	14	236,532			8
ARS-BFGL-NGS-4939	14	443,937		✓	10
ARS-BFGL-NGS-107379	14	679,600			10
BFGL-NGS-113453	14	30,002,363	✓		9
Hapmap48989-BTA-101611	14	34,879,141			9
BTB-01157350	17	1,384,889			9
BTB-01951543	20	49,918,496			8
Hapmap39714-BTA-111678	21	18,811,723			8
ARS-BFGL-NGS-69616	21	22,250,027			9
Protein (%)					
Hapmap39813-BTA-21834	1	53,621,500			8
BTB-01978832	2	135,640,997			8
Hapmap42708-BTA-86534	3	50,850,297	✓		9
Hapmap50895-BTA-122111	4	24,553,482			10
BTB-00234759	5	94,104,074			8
BTB-01534149	6	66,230,967	✓		9
ARS-BFGL-NGS-29273	7	5,652,920			8
BTA-87610-no-rs	7	57,673,607			9
ARS-BFGL-NGS-64882	7	81,392,639			10
Hapmap49034-BTA-115720	8	88,958,729			8
Hapmap39516-BTA-82096	8	90,551,290			8
ARS-BFGL-NGS-17358	8	102,312,736			9
ARS-BFGL-NGS-100341	9	52,911,776	✓		8
BTA-63354-no-rs	10	33,150,420			9
BTB-01972463	11	16,078,669			9
BTA-93103-no-rs	11	33,330,852			8
ARS-BFGL-NGS-32722	11	95,942,521			10
Hapmap58072-rs29010006	12	63,149,779			9
BTB-01123944	13	6,281,252			8
ARS-BFGL-BAC-11928	13	29,191,010			8
BTA-32552-no-rs	13	42,633,511			9
ARS-BFGL-NGS-38064	13	54,700,987			8
ARS-BFGL-NGS-104967	13	55,847,196			9
ARS-BFGL-NGS-5166	13	56,045,155			9
ARS-BFGL-NGS-57820	14	236,532			10
ARS-BFGL-NGS-4939	14	443,937		✓	10
ARS-BFGL-NGS-107379	14	679,600			10
Hapmap30086-BTC-002066	14	1,490,178			9
BFGL-NGS-113453	14	30,002,363			9
Hapmap41433-BTA-114994	14	33,550,219			9

SNP marker	Chromosome	Position in bp	QTL region	QTL peak	Frequency
Hapmap48989-BTA-101611	14	34,879,141	✓		9
BTB-00642563	16	43,274,808			9
BTA-40059-no-rs	16	694,769,20			9
BTB-01157350	17	1,384,889			9
Hapmap49611-BTA-44077	17	26,304,955			8
Hapmap42359-BTA-90829	18	20,737,022			9
Hapmap49176-BTA-43744	18	49,761,247			10
ARS-BFGL-NGS-34276	18	49,839,669			9
Hapmap34814-BES8_Contig361_961	19	20,361,224			8
ARS-BFGL-NGS-11174	19	43,331,499			8
ARS-BFGL-NGS-69616	21	22,250,027			9
Hapmap49032-BTA-115439	22	54,473,771			9
BTA-54892-no-rs	22	54,503,230			8
BTA-112061-no-rs	23	38,207,532			8
ARS-BFGL-NGS-22050	25	27,843,968			10
ARS-BFGL-BAC-42500	25	28,002,712			10
ARS-BFGL-NGS-41056	26	20,364,191			9
BTB-00624015	27	20,648,605			10
ARS-BFGL-NGS-18177	29	4,598,272			10
ARS-BFGL-NGS-20615	29	5,294,603			8
pH value					
ARS-BFGL-NGS-103495	8	9,605,960			9
BTA-63354-no-rs	10	33,150,420			10
BTA-98790-no-rs	13	25,929,330			9
ARS-BFGL-BAC-12549	13	54,601,927			9
ARS-BFGL-NGS-57820	14	236,532			8
ARS-BFGL-NGS-4939	14	443,937			9
ARS-BFGL-NGS-107379	14	679,600			10
ARS-BFGL-NGS-107810	15	66,121,820			8
ARS-BFGL-NGS-40131	17	10,374,164			8
Hapmap49611-BTA-44077	17	26,304,955			9
Hapmap49176-BTA-43744	18	49,761,247			10
BTA-23545-no-rs	18	50,063,553			8
ARS-BFGL-NGS-22050	25	27,843,968			10
ARS-BFGL-BAC-42500	25	28,002,712			10
ARS-BFGL-NGS-100347	25	28,535,691			9
BTB-00624015	27	20,648,605			10
Hapmap42281-BTA-63982	28	29,215,768			8
ARS-BFGL-NGS-18177	29	4,598,272			8

✓- SNP is located in a QTL

B.7 Investigations of the importance of DGAT1 on three selected milk traits.

In an additional study the impact of DGAT1 was investigated on fat content, protein content and pH value. For this study the following settings were used:

1. All SNPs were used.
2. DGAT1-SNP (SNP marker: ARS-BFGL-NGS-4939) was excluded from the SNP set.
3. DGAT1-region was excluded from the SNP set, wherein the DGAT1-region was defined from 0 to 3,940,998 bp on chromosome 14 (85 SNPs were excluded in total). The end position was chosen based on the right marker position of the QTL with ID 3172 (CSSM066; cf. Appendix B.3.)

The analyses were realized using SVS (settings as in Section 4.2.4 on page 83) and based on the 10-fold cross-validation design (cf. Section 4.2.2 on page 77).
In addition, the DGAT1-region was also analyzed, since it is known that neighboring SNPs can even capture the genetic effect if the main SNP in this region is not available. In Table B.5 the results of the genetic value prediction of this study are presented. In this table can be seen that if only the DAGT1-SNP is excluded then the mean prediction precisions are similar to the results containing the DAGT1-SNP. In contrast, if the DAGT1-region is excluded then we observed that the prediction precision significantly decreases for all investigated milk traits: 47.16% for fat content, 29.96% for protein content and 22.15% for pH value.

Table B.5: The mean prediction precisions are listed for all SNPs, without DGAT1-SNP and without DGAT1-region for all three investigated milk traits. In parentheses the corresponding standard deviations are presented.

Milk trait	All SNPs		without DGAT1-SNP		without DGAT1-region	
Fat (%)	0.299	(0.077)	0.278	(0.077)	0.158	(0.077)
Protein (%)	0.237	(0.056)	0.231	(0.057)	0.166	(0.069)
pH value	0.307	(0.090)	0.302	(0.088)	0.239	(0.087)

Acknowledgement

First of all I would like to thank my supervisor, Dr. D. Repsilber, who gave me the possibility to realize this thesis and especially for the confidence that he placed in me. During the whole time he supported me and gave me not only useful advice for my research but also encouraged me to improve my own skills.

Thanks to Prof. O. Wolkenhauer for mentoring this work and to his research group where I also got helpful tips and suggestions.

Many thanks to Dr. D. Wittenburg, Dr. S. Trißl, Dr. S. Andorf and Dr. D. Zimmer for all their help, support and patience during the project. In addition, I would like to thank for their useful suggestions and comments while I wrote this thesis. Here a special thanks to Dr. C. Baes and Dr. S. Trißl. Furthermore, I would like to thank all present and former members of the Institute of Genetics and Biometry (Leibniz Institute for Farm Animal Biology, FBN, Germany) for their advice and support performing this thesis.

In the following I would like, also on behalf of the research group, to thank all the people which were involved and assisted us during data collection, preparation or post-preparation and especially for the good cooperation. Without these people the experimental data would not exist.

- A special thanks goes to the 18 farms within Mecklenburg-Western Pomerania (Germany) for their cooperation and for giving us the possibility to realize this project.
- Dr. F. Reinhardt and E. Pasmann (VIT Verden, Germany) who provided us the pedigree information and also all information of the dams.
- Dr. U. Kesting, Dr. S. Jakubowski, Dr. S. Hartwig and S. Wolf (LKV, Güstrow, Germany) managed and controlled all around the milk traits collection and provided us with the measured milk traits information. They made sure that our milk samples were measured always together, using the same machine. An additional thanks to the LKV milk performance inspectors.
- The team around Prof. Dr. T. Meitinger (HelmHoltz Zentrum Munich, Germany) prepared the SNP-genotypes from the proven DNA-samples.
- Prof. L. Willmitzer, Dr. J. Lisec and Ä. Eckardt (Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany) were involved for the milk metabolite

profiling. A special thank goes to Ä. Eckardt, because she has established and validated the measuring of the milk metabolites for the first-time in their lab. In addition, she has measured the main part of our milk samples.

- PD Dr. J. Vanselow, Dr. R. Fürbaß, M. Nimz and M. Anders, M. Spitschak (FBN, Institute of Reproductive Biology) who provided us facility and gave me useful advices if I had questions.
- Dr. H. Hammon, C. Reiko (FBN, Institute of Nutritional Physiology "Oskar Kellner") who provided us facilities to fill the taken blood samples into the tubes.
- PD Dr. C. Kühn and S. Wöhl (FBN, Institute of Genome Biology) realized, together with us, the test run with the used NucleoSpin BloodL toolkit (Machery-Nagel, Düren, Germany).
- A special thank goes to R. Grahl (FBN, Institute of Genetics and Biometry), who made appointments and visited the farms as well as to the veterinarians for collecting the blood samples. He also made appointments with the LKV Güstrow to collect the milk samples.
- A. Rief (FBN) who created the marker map for us.

Finally, I thank the German Federal Ministry of Education and Research (BMBF) for the funding of the Fugato Plus project BovIBI (Bovine Integrative BioInformatics for genomic selection; funding code: 0315137) and the Deutsche Forschungsgemeinschaft (DFG) for the funding of the project analysis of microRNAs involved in malignant melanoma progression with the funding code: GZ:WO0991/4-1.

Special thank to M. Hellmig who supported me during the whole time, especially when I had lost my smile he cheered me up. Thanks a lot.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und unter Verwendung keiner anderen als den von mir angegebenen Quellen und Hilfsmitteln verfasst habe, sowie Zitate kenntlich gemacht.
Ferner erkläre ich, dass ich bisher weder an der Universität Rostock noch anderweitig versucht habe, eine Dissertation einzureichen oder mich einer Doktorprüfung zu unterziehen.

Nina Melzer
Rostock, den 17.05.2013

Publications and own contributions

The following publications have resulted from this thesis:

Journal publications - peer reviewed

1. N. Melzer, D. Wittenburg, S. Hartwig, S. Jakubowski, U. Kesting, L. Willmitzer, J. Lisec, N. Reinsch and D. Repsilber (2013a). *Investigating associations between milk metabolite profiles and milk traits of Holstein cows.* Journal of Dairy Science. 96(3):1521-1534. doi:10.7482/0003-9438-56-038.

The study was conceived by D. Repsilber and N. Reinsch. N. Melzer was mainly involved in data collection and preparation of the recorded milk traits and used SNP-genotypes (cf. Section "Own contribution regarding the experimental data collection and preparation").
The analyses were performed and the interpretation of the results were realized by N. Melzer and supervised by D. Repsilber. N. Melzer performed the literature search and wrote main parts of the manuscript with assistance of D. Wittenburg and D. Repsilber. S. Hartwig, S. Jakubowski and U. Kesting were involved in milk sampling and gave advice to handle the milk data. They read and gave comments on the manuscript. L. Willmitzer and J. Lisec were involved in metabolite profiling and the post-processing steps involved. They read and gave comments on the manuscript. N. Reinsch read and also gave comments on the manuscript. N. Melzer was mainly involved in answering the reviewer comments.

Journal of Dairy Science Impact $\text{Factor}_{2012} = 2.56$
Journal of dairy science has the rank two out of 55 journals in total in the field Agriculture, Dairy and Animal Science (accessed 24.04.2013).

2. N. Melzer, D. Wittenburg and D. Repsilber (2013b). *Investigating a complex genotype-phenotype map for development of methods to predict genetic values based on genome-wide marker data - a simulation study for the livestock perspective.* Archiv Tierzucht, 56(38):380398.

D. Repsilber raised the initial question. The concept of how to realize the simulated metabolome level within the GP map was designed in collaboration between N. Melzer and D. Repsilber. N. Melzer simulated the data and also realized the necessary preliminary analysis to enable the simulation of populations with a more or less realistic LD based on an available SNP chip annotation (Melzer et *al.* 2010b, No. 9). All analyses were performed by N. Melzer. The interpretation of the results was realized by N. Melzer with support of D. Wittenburg and D. Repsilber. D. Wittenburg contributed the extended estimation method of the fast BayesB algorithm as well as contributed to the analyses. N. Melzer was mainly involved in data collection and preparation of the recorded milk traits and SNP-genotypes (cf. Section "Own contribution

at the experimental data collection and preparation"). N. Melzer performed the literature search and wrote main parts of the manuscript with assistance of D. Wittenburg and D. Repsilber. N. Melzer was mainly involved in answering the reviewer comments.

Archiv für Tierzucht: Impact Factor$_{2011}$ = 0.42

3. N. Melzer, D. Wittenburg and D. Repsilber (2013c). Analyzing milk metabolite profiles to enable prediction of traditional milk traits for Holstein cows based on SNP information. PLoS ONE, 8(8):e70256. (This study was under review for publication at thesis submission.)

D. Repsilber gave the basic idea. The corresponding preliminary study to test our proposed metabolite approach on simulated data to verify if and to which degree an improvement of the genetic value prediction can be achieved in respect to experimental data was born as the three system-levels were simulated (Melzer et al. 2011, No. 10). The concept of the analyses as well as validation proposals were developed by N. Melzer and was supervised by D. Repsilber. The interpretation of the results was realized by N. Melzer and was supervised by D. Repsilber. D. Wittenburg contributed the extended stochastic variable selection method, which was used for genotype-phenotype prediction as well as SNP selection. N. Melzer performed the literature search. N. Melzer wrote the manuscript and was supervised by D. Repsilber and D. Wittenburg.

Journal of Dairy Science Impact Factor$_{2012}$ = 2.56

Conference contributions - peer reviewed

4. N. Melzer, S. Jakubowski., S. Hartwig, U. Kesting, S. Wolf, F. Reinhardt, E. Pasman, G. Nürnberg, N. Reinsch and D. Repsilber (2010a). *Design, infrastructure and database structure for a study on predicting milk phenotypes from genome wide SNP markers and metabolite profiles.* In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Abstract ID 0427. Leipzig, Germany. ISBN:978-3-00-031608-1.

N. Melzer created the database and was mainly involved in data collection and preparation of the data, more information are provided in Section "Own contribution regarding the experimental data collection and preparation". The milk randomization design was created in cooperation between G. Nürnberg, D. Repsilber and N. Melzer. The implementation via an R-script was realized by N. Melzer. G. Nürnberg read and gave comments on the manuscript. S. Hartwig, S. Jakubowski, U. Kesting and S. Wolf were involved in milk sampling and gave advice on handling milk data. They read and commented on the manuscript. F. Reinhardt and E. Pasman provided the pedigree information and also all information of dams. Both read and gave comments on the manuscript. N. Reinsch was mainly involved to make arrangements with the partners. N. Reinsch read the manuscript and gave comments. N. Melzer wrote the manuscript and was supervised by D. Repsilber.

"This congress is the premier conference for researchers and professionals involved in genetic improvement of livestock. Delegates from around the world gather every four years to attend the scientific program and network with colleagues." (adapted from the official homepage

"wcgalp.com", accessed on 27.04.2013). 1,370 delegates from 59 nations participated the 9^{th} WCGALP in Leipzig.

5. N. Melzer, D. Wittenburg and D. Repsilber (2012). *Metabolites as new molecular traits and their role for genetic evaluation of traditional milk traits.* In: 63rd Annual Meeting of the EAAP. Bratislava, Slovakia. ISSN:1382-6077. p. 88. Doi: 10.3920/978-90-8686-761-5.
EAAP scholarship holder (13 scholarships/48 requests)

The basic idea gave D. Repsilber. The corresponding preliminary study to test our proposed metabolite approach on simulated data to verify if and to which degree an improvement of the genetic value prediction can be achieved in respect to experimental data was born as the three system-levels were simulated (Melzer et al. 2011, No. 10). The concept of the analyses proposals were developed by N. Melzer and was supervised by D. Repsilber. D. Wittenburg contributed the extended stochastic variable selection method, which was used for genotype-phenotype prediction as well as for SNP selection. N. Melzer realized all analyses and interpreted the results. The latter was supervised by D. Repsilber. N. Melzer performed literature search. N. Melzer wrote the manuscript and was supervised by D. Repsilber and D. Wittenburg.

"One of the worldwide most important scientific meetings of professionals working in the area of animal production." (adapted from the "Book of Abstracts of the 63^{rd} Annual meeting of the European Federation of Animal Science".)

Co-authored publications - peer reviewed

Journal publications

6. D. Wittenburg, **N. Melzer** and N. Reinsch (2011). *Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers.* BMC Genetics 12(74).

N. Melzer generated the simulated data sets and contributed to the data analysis. N. Melzer read and improved the manuscript.

BMC genetics: Impact $\text{Factor}_{2012} = 2.48$

7. D. Wittenburg, **N. Melzer**, L. Willmitzer, J. Lisec, U. Kesting, N. Reinsch and D. Repsilber (2013). *Milk metabolites and their genetic variability.* Journal of Dairy Science. 96(4), 2257-2569.

N. Melzer was mainly involved in data collecting and contribute to post-processing of the obtained milk metabolite spectra from the Max Planck Institute of Molecular Plant Physiology (Potsdam-Golm, Germany; cf. Section "Own contribution regarding the experimental data collection and preparation"). N. Melzer assigned milk metabolites to their chemical classes. N. Melzer contributed to the the manuscript.

Journal of Dairy Science Impact $\text{Factor}_{2012} = 2.56$

Conference contribution

8. D. Wittenburg, **N. Melzer** and N. Reinsch (2010a). *Including non-additive effects in Bayesian methods for the prediction of genetic values from genome-wide SNP data.* In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production (WCGALP). Leipzig, Germany. Abstract ID 0267, ISBN:978-3-00-031608-1.

Publications - non-peer reviewed

9. N. Melzer, D. Wittenburg and D. Repsilber. (2010). *Simulating SNP data: influence of simulation design on the extent of linkage disequilibrium.* In: H.-M. Seyfert and G. Viereck (editor), 11th Day of the Doctoral Student (FBN Dummerstorf), p. 19-22. BUK! Breitschuh & Kock, Kiel, Germany. ISSN:0946-1981.

The concept of the analyses was mainly developed by N. Melzer and was supervised by D. Wittenburg and D. Repsilber. All analyses were realized and the obtained results were interpreted by N. Melzer. The latter was supervised by D. Wittenburg and D. Repsilber. The paper was written by N. Melzer with assistence of D. Wittenburg and D. Repsilber.

10. N. Melzer, D. Wittenburg and D. Repsilber. (2011). *Including metabolomic profiles to improve genetic value prediction: an integrated bioinformatics approach using weighted genome-wide marker information.* In: H.-M. Seyfert and G. Viereck (editor), 12th Day of the Doctoral Student (FBN Dummerstorf), p. 55-58. BUK! Breitschuh & Kock, Kiel, Germany. ISSN:0946-1981.

The concept of the analysis was mainly designed by N. Melzer and was supervised by D. Wittenburg and D. Repsilber. The analyses were performed by N. Melzer. Results were interpreted by N. Melzer which was supervised by D. Wittenburg and D. Repsilber. N. Melzer wrote the paper with assistance of D. Repsilber and D. Wittenburg.

Presentations by N. Melzer

11. Including metabolomic profiles to improve genetic value prediction: an integrated bioinformatics approach using weighted genome-wide marker information. 12th Day of the Doctoral Student 2010 (FBN Dummerstorf), Germany.

12. Accounting for a complex genotype-phenotype map in milk phenotypes from genome-wide marker data. In: Statistical Computings 2010 (Reisensburg), Germany.

13. Simulating SNP data: influence of simulation design on the extent of linkage disequilibrium. 11th Day of the Doctoral Student (FBN Dummerstorf), Germany.

Posters at international conferences

14. N. Melzer, D. Wittenburg and D. Repsilber (2009). *A simulation approach for genotype-phenotype-mapping via metabolome.* GCB 2009, German Conference on Bioinformatics, Halle (Saale), Germany.

15. N. Melzer, D. Wittenburg and D. Repsilber (2010). *Simulating a Complex Genotype-phenotype Map for Development of Genomic Selection Methods.* ICSB 2010: International Conference on Systems Biology, Edinburgh, Scotland.

16. D. Repsilber, D. Wittenburg and **N. Melzer** (2011). *Integrating metabolome information for prediction of milk phenotypes from genome-wide SNP data.* ICSB2011: International Conference on Systems Biology, Heidelberg/Mannheim, Germany.

17. D. Wittenburg, **N. Melzer**, N. Reinsch and D. Repsilber (2012). *Milk metabolites and their genetic variability.* In: Book of Abstracts of the 63rd Annual Meeting of the EAAP, 95. Wageningen Academic Publishers, Bratislava, Slovakia. ISSN:1382-6077.

Own contribution regarding the experimental data collection and preparation

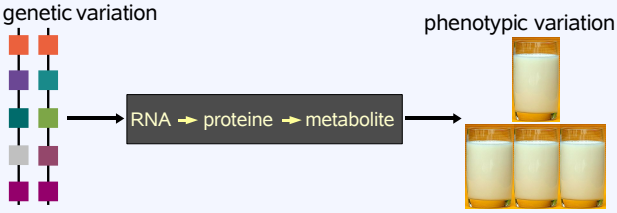
The following section was supervised by D. Repsilber. N. Melzer curated and validated the incoming various information from the co-operation partners, populated the database, and ensured the sample allocation as well as monitored the number of the corresponding samples over the entire sample collection. The presented BovIBI database (Chapter 2) was created, maintained and verified by N. Melzer. The corresponding R-scripts, including mySQL statements, were written by N. Melzer. The weekly updated milk lists were mainly generated and provided for the LKV Güstrow by N. Melzer. The blood samples, after a test run with a technician, were prepared by N. Melzer as well as the corresponding DNA quality and quantity checks. The milk samples were prepared and the corresponding guidance for Max Planck Institute of Molecular Plant Physiology (Potsdam-Golm, Germany) was realized by N. Melzer. The post-processing steps of the obtained milk metabolite spectra from the Max Planck Institute of Molecular Plant Physiology were mainly realized by D. Repsilber in co-operation with J. Lisec (Max Planck Institute of Molecular Plant Physiology). N. Melzer assisted in these steps. R. Grahl collected the milk samples from the LKV Güstrow, as well D. Repsilber and N. Melzer.

A simulation approach for genotype-phenotype-mapping via metabolome

Nina Melzer, Dörte Wittenburg and Dirk Repsilber

Motivation and Problem

genetic variation

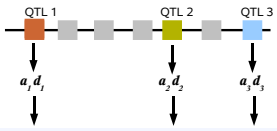


genomic selection aims to predict the phenotype using the information from genome-wide marker data

new: a more realistic simulation approach using the concrete metabolome level

Statistical Model

conventional approach



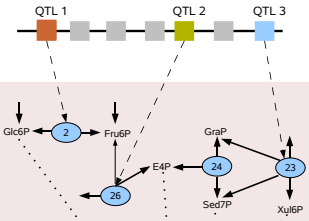
Genotypes
52273 SNPs
$a_i \sim \Gamma(0.42, 0.382)$ – allele substitution effect[1,2]
$d_i = k_i \mid a_i \mid$ – dominance effect
$k_i \sim N(0.193, 0.312^2)$[3]

$$\sum_{i=1}^{\#loci} \left(X_{ji} a_i + D_{ji} d_i \right) + e_j$$

X_{ji} – design matrix with genotypes
D_{ji} – incidence matrix for heterozygosity
$e_j \sim N(0, \sigma_e^2)$
j – animal index

Phenotype

SBML approach



Genotypes
52273 SNPs

The enzyme parameters change according to the genotype[4]. The heterozygotes get the default value and the homozygotes ± 50%. (part of the BIOMD70[5])

$$\sum_{i=1}^{\#loci} \sum_{l=1}^{\#products} P_{jil} + f_j$$

P_{jil} - enzyme end concentrations
$f_j \sim N(0, \sigma_f^2)$

Phenotype

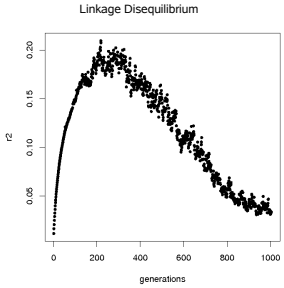
Simulated Data

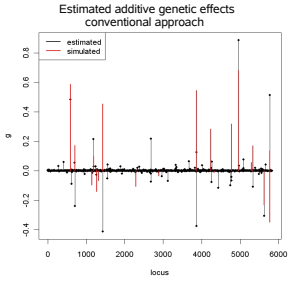
Population structure	N	Marker genotyping	Phenotypic recording	
Generations 1-200	100	-	-	
Generation 201	1000	yes	yes	training
Generation 202	1000	yes	yes	training
Generation 203	1000	yes	-	test

Conclusions

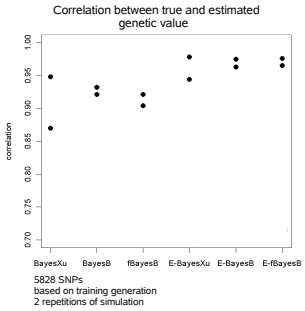
- The first results show that conventional Bayesian methods can be used to learn and predict phenotypes from genome-wide genetic marker data using the more realistic SBML approach.
- Set-up of simulation affects the extent and permanence of the Linkage Disequilibrium.

Results

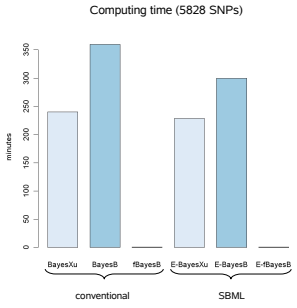




- Fast Bayes required a split of computing time compared to other Bayesian methods (no Gibbs-Sampler involved).
- The result of fast Bayes gives a trend which is confirmed by the Bayes B.
- Bayes B gives a higher precision than Xu's Bayes as expected from the literature.
- Reparameterisation of additive and dominance effects is necessary to avoid covariances between those.
- The choice of reparameterisation has an influence on the precision of parameter estimation.



5828 SNPs
based on training generation
2 repetitions of simulation



Outlook

- consideration of interactions (epistatic effects) in Bayesian analysis
- consider mutation of alleles when simulating the generations
- generate further data sets for validation
- estimate the proportion of QTL to SNP markers

References

[1] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. Genetics,157(4):1819–29, 2001.
[2] B. Hayes and M. E. Goddard. The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol, 33(3):209–29, 2001.
[3] J.Bennewitz and T.H. Meuwissen,The distribution of additive and dominant QTL effects in porcine F2 crosses, Book of Abstracts of the 60th Annual Meeting of the EAAP, Barcelona, Spain (2009),p.320
[4] P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics, 19 Suppl 2:ii122–9, 2003.
[5] http://www.ebi.ac.uk/biomodels-main/

Research Institute for the Biology of Farm Animals, Genetics & Biometry
Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany
http://www1.fbn-dummerstorf.de/de/Forschung/FBs/fb2/repsilber/AG/AG.html

GCB2009, Halle (Saale), September 28-30, 2009

Leibniz Gemeinschaft

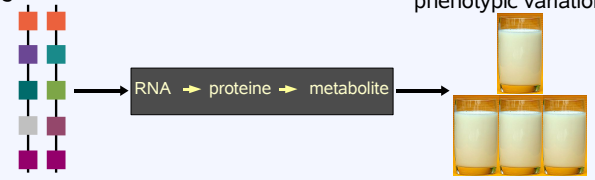
Simulating a Complex Genotype-phenotype Map for Development of Genomic Selection Methods

FBN / Wilhelm-Stahl-Allee 2 / 18196 Dummerstorf / www.fbn-dummerstorf.de

Nina Melzer, Dörte Wittenburg, Dirk Repsilber

Motivation and Problem

genetic variation → RNA → proteine → metabolite → phenotypic variation

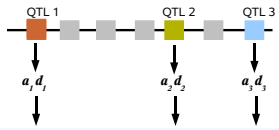


genomic selection aims to predict the phenotype using the information from genome-wide marker data

new: a more realistic simulation approach using the concrete metabolome level

Statistical Model

conventional approach



Genotypes
52273 SNPs

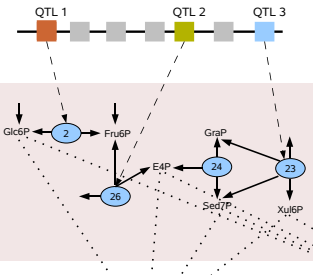
$a_i \sim \Gamma(0.42, 0.382)$ - allele substitution effect[1,2]
$d_i = k_i |a_i|$ - dominance effect[3]
$k_i \sim N(0.193, 0.312^2)$[5]

$$\sum_{i=1}^{\#loci} X_{ji} a_i + D_{ji} d_i + e_j$$

X_{ji} – design matrix with genotypes
D_{ji} – incidence matrix for heterozygosity
$e_j \sim N(0, \sigma_e^2)$
j - animal index

↓ **Phenotype**

SBML approach



Genotypes
52273 SNPs

The enzyme parameters change according to the genotype[3]. The heterozygotes get the default value and the homozygotes ± 50%. (part of the BIOMD70[4])

SBML 1

$$\sum_{i=1}^{\#loci} \sum_{l}^{\#products} P_{ijl} + f_j$$

SBML 2

$$\sum_{i=1}^{\#loci} \sum_{l}^{\#products} \sin(P_{ijl}) + f_j$$

P_{ijl} - enzyme end concentrations
$f_j \sim N(0, \sigma_f^2)$

↓ **Phenotype** ↓ **Phenotype**

Set-up

Simulation:

- 100 data set (mutation-drift model)
- 2 training and 2 test generations (pro generation 1000 animals)
- quantity of SNPs: 5227, 52273
- quantity of QTL: 23, 230

Evaluation

- γ: proportion of non zero effects
- Preliminary analysis: fastBayes in terms of γ -> focus on estimation γ
- γ = {0.1, 0.5, 0.25, 0.01,0.005, 0.001, 0.0001, 0.00001}

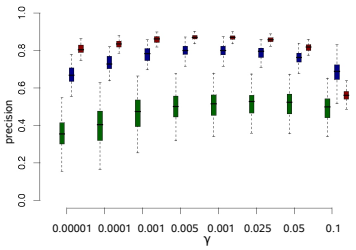
} all possibilities were analyzed for conventional, SBML 1, SBML 2

Conclusions

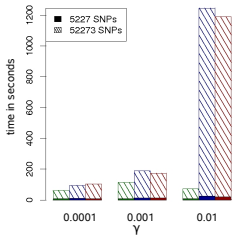
- The results show that the conventional fastBayes method can be used to learn and predict phenotypes from genome-wide genetic marker data using the more realistic SBML approach.
- Further it can be concluded that a linear simulated genotype-phenotype map allows for higher precision predictions than a non linear mapping.

Results

predicting precision comparison for SBML 1, 230 QTLs and 5227 SNPs

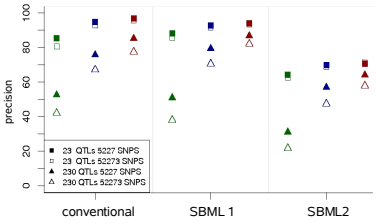


computing time comparison for SBML 1 and 230 QTLs

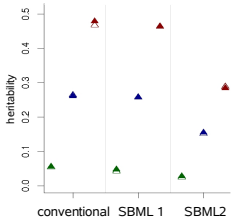


- The reduced data set (5227 SNPs) shows similar results regarding precision and heritability estimation as for using the complete 52273 SNPs data set.
- Conventional and SBML approach show similar behavior in view of precision and different γ.
- Estimated heritabilities are similar for all possibilities for each approach.

comparison best predicting precision for all approaches



estimated heritability for γ with best predicting precision



approach		simulated proportions of effects (γ)		best precision values of γ	
method	#QTLs	5227 SNPs	52273 SNPs	5227 SNPs	52273 SNPs
conventional, SBML 1, SBML 2	23	0.0044	0.00044	0.001	0.0001
conventional, SBML 1, SBML 2	230	0.044	0.0044	0.01	0.001

legend: ■ = heritability 0.1 ■ = heritability 0.3 ■ = heritability 0.5

Outlook

- consideration of two fold interactions in Bayesian analysis (epistatic effects)
- using simulated metabolite data for enhanced phenotype prediction
- weighting SNPs
- validation with experimental data

References

[1] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. Genetics,157(4):1819–29, 2001.
[2] B. Hayes and M. E. Goddard. The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol, 33(3):209–29, 2001.
[3] P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective com parison of analysis algorithms. Bioinformatics, 19 Suppl 2:ii122–9, 2003.
[4] http://www.ebi.ac.uk/biomodels-main/
[5] J.Bennewitz and T.H. Meuwissen,The distribution of additive and dominant QTL effects in porcine F2 crosses, Book of Abstracts of the 60th Annual Meeting of the EAAP, Barcelona, Spain (2009),p.320

ICSB 2010, Edinburgh, October 10-15, 2010

LEIBNIZ INSTITUTE FOR FARM ANIMAL BIOLOGY

correspondence
melzer@fbn-dummerstorf.de

Theses

Nina Melzer, Investigating possibilities to predict milk phenotypes in *Holstein Friesian* cows based on a more complex model of the genotype-phenotype map

An alternative (SBML) approach is proposed to simulate more realistic data, in respect to collected experimental data, including the simulation of genotype, phenotype (as in the conventional approach) and metabolome data. To enable analysis of these system-levels an integrative bioinformatics approach (metabolite approach) is proposed, which contains the following steps: metabolome-phenotype, genotype-metabolome and genotype-phenotype:

1. The SBML approach used a systems biology approach, i.e., metabolic network model, to simulate metabolome data depending on the genotype. A simple additive step was used to simulate phenotypes based on the metabolome. In contrast to the conventional approach, the SBML approach enables to simulate genetic effects implicitly by the interactions of the used metabolic network model.

2. Investigations of the implicitly simulated genetic effects revealed that some simulated QTL had no impact on their metabolic outcomes (thus on the phenotype) within the used metabolic network.

3. Data simulated with the SBML approach enable various further methodological investigations as well as to test different analysis possibilities, in contrast to data simulated with the conventional approach.

4. The metabolite approach was applied on simulated data to investigate the gain of using the metabolome level as an additional information source for the genetic value prediction. Our results revealed that it is possible to improve the genetic value prediction using the additional information. The success of improving depends on the used simulated part of the metabolome.

5. The prediction of milk fat content (1.300 cows) revealed comparable prediction precisions obtained with the metabolite approach (129 SNPs) and the classical approach (40,317 SNPs), wherein the metabolite approach required only fraction of the total amount of SNPs.

6. Significantly more selected SNPs, using a variable selection method, were found on known QTL via the metabolite approach than for using the classical approach, for fat content and protein content. New important genome regions can possibly be revealed for an investigated milk trait.

7. Results regarding the number of selected SNPs, point towards less complexity of the underlying genetic structure of most of milk metabolites compared to milk traits.

8. The milk metabolome-phenotype map was investigated in greater detail using uni- and multivariate methods, where new relations were revealed.