Design, Infrastructure And Database Structure For A Study On Predicting Of Milk Phenotypes From Genome-Wide SNP Markers And Metabolite Profiles

N. Melzer[*], S. Jakubowski[†], S. Hartwig[†], U. Kesting[†], S. Wolf[†], G. Nürnberg[*], N. Reinsch[*] and D. Repsilber[*]

Introduction

We aim to compare two different approaches to predict milk phenotypes from cows using genome-wide marker data. The conventional genotype-phenotype-mapping approach (e.g. Meuwissen et al. (2001)) should be compared to our alternative proposal that includes information about the metabolome level. The objective is to find out if collecting additional molecular information on the genotype-phenotype-mapping will improve phenotype prediction. A simulation study is under way, but its results are to be validated using experimental data. This contribution is to survey obstacles, key organization, infrastructure and design issues, as well as our database approach to secure collection of valid experimental data.
To reach the three levels of genotype, metabolome and phenotype, it is required to have both blood and milk samples of the selected 1,300 cows. With a sample of blood it is possible to access the genotype after DNA-extraction and SNP genotyping. The sample of milk can be used for the milk performance test (MLP), where different milk features are measured like fat, protein and quantity of milk, as phenotypes. Additionally, the sample of milk can be used to access the metabolome level by using a gas chromatography mass spectrometry (GC-MS) approach (Lisec, J., Schauer, N., Kopka, J. et al. (2006)). Our experimental set-up implies massive amounts of data and various information from involved co-operating partners. To handle and organize all information, as well as to secure data integrity and validity, a database was built.

Material and methods

Data collection. We aim to measure blood and milk samples from 1,300 cows. Each cow has to fulfill the following conditions: first, the cow must have its first lactation and second, the sample of milk must be taken between the 20th and 120th day of lactation. In summa, 1,775 cows from 18 agriculture holdings within Western Pomerania were selected. We made arrangements with the agricultural holdings and the regional institute for standard milk performance testing (Landeskontrollverband für Leistungs- und Qualitätsprüfung, LKV Güstrow,

[*]FBN Dummerstorf, Wilhelm-Stahl-Allee 2, Dummerstorf, Germany
[†]LKV, Güstrow, Germany

Germany). We received ear tag numbers from the agricultural holdings which were sent to the institute IT-solutions for animals (VIT Verden, Germany; we thank Dr. F. Reinhardt and E. Pasman for their collaboration). From the VIT we got pedigree information and date of calving for each cow. Afterwards we could start with sampling of blood (period: January-February and October-November 2009). During the period of milk sampling (March-November 2009) we received weekly updates on which cows were still alive, had calved and the day of calving, for all selected cows, from VIT. We used this list to generate a cow-list for every agricultural holding for taking the milk samples and provided the different lists for the LKV Güstrow as regularly updates websites. Monthly, we received from the LKV zip-files with information of measured values from the MLP for the milk samples. Additionally, we had to evaluate milk lists coming back after being used by the performance inspectors, with annotations about cows which were sold, medical treatments or disease status.
During the period of data collection we had to cope with processing various types of information, changing formats, missing values, errors and unforeseeable events. After completion, we have in total 1,342 cows for which we got all types of measurements are complete. The data collection was done during standard business in commercial herds.

Problems. The following conditions have an influence of the milk phenotype: time point of taking milk sample, agriculture holding (feeding, type of milk collection), and from the genetic side the pedigree information, especially knowledge about groups of half-sibs.

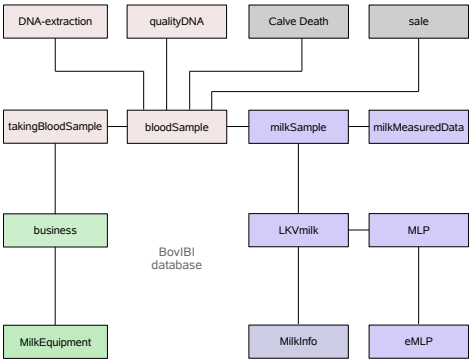
Questions. How to organize all the different information from the data collection? How to consider the influencing factors above during the measuring of metabolite profiles?

Results and discussion

Database. During the data collection a MySQL[1] database, shown in figure 1, was built and expanded according to the additional new information. Furthermore, we used phpMyAdmin, a free software tool, to handle the administration of mySQL in a graphical user interface, and share all information within the working group. The database is an instrument to monitor, allow common access and facilitate handling and checking the available information. The database simplifies the processing of the data, and further it is possible to connect to the database from other programming and analysis software like R (R Development Core Team, 2005). For example, we used the R package RMySQL (James and DebRoy, 2004) to prepare the milk lists and websites for the LKV Güstrow. Moreover, the database is based on tables, structured into four parts. One part is for the information of the agriculture holdings and the corresponding cows and samples of blood (figure 1, green). The next part is for the DNA-extraction and SNP-chip information (figure1, red). The third part contains information of lost and sold cows (figure 1, grey) and the fourth part all information about the samples of milk and milk metabolite profiles (figure 1, blue). We got several information at different time points. Throughout the whole data collection it was necessary to know how much data in each part of the database are complete or erroneous and how many cows have all desired information complete. This was necessary to reach the desired number of 1,300 cows. This information were used to design a second collection of blood samples. Without a database it

[1] MySQL: relational database management system

is a hard task to have an overview over the different kinds of data and to find errors or make plausibility checks. The full database structure is available upon request to the authors.



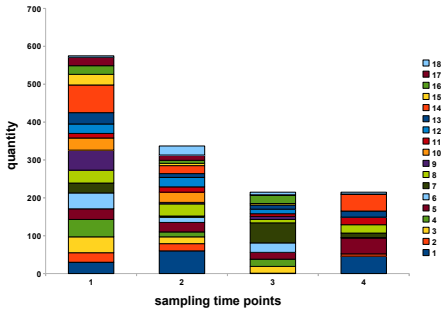


Figure 1: (left) Part of the BovIBI database
Figure 2: (right) Overview of the quantity of milk samples of the coarsened time points from the 18 agricultural holdings

Table 1: Part of the milk metabolite measuring design

	measure day 1			measure day 2			measure day 3		
holding	**holding**	**sampling**	**sire**	**holding**	**sampling**	**sire**	**holding**	**sampling**	**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

Randomization for milk metabolite profiling. We have 1,342 labelled samples of milk and for each sample the following information: ear tag number from each cow and the corresponding sire. In total, we have 7 time points (March till November 2009) from milk sampling. The desired design for the conditional randomization should fulfill the following criteria:

1. each agriculture holding occurs at each day of measuring
2. two sampling time points at each day of measuring
3. half-sibs should be measured on consecutive days of measuring

The design was based on the latin square design. The order of the conditions represents their importance ranking – to be considered for this randomization.
It is possible to measure 40 samples in GC-MS per day and we had to plan for milk metabolite profiling on 34 days. The quantity of samples is imbalanced over the months and

varies strong (not shown here). The frequency of the sires varies strong (in total 216 different sires). To get a better result for criterion 3, the 7 sampling time points are coarsened into 4 time points:

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

All conditions are so far considered as far as possible. That means we tried to produce a GC-MS design table balanced as possible. Part of it is shown in figure 2, here the sires there are coloured to show that half-sibs are measured on consecutive days. The full design is available upon request.

Conclusion

Our experiences and results show that it is necessary to always have an overview of the data during collection, and that a database is good and useful instrument to handle and organize data in various respects. Furthermore, phpMyAdmin is a user friendly administrative tool for database. The database simplifies the access to the data and makes it easy to get various information during ongoing data collection. This way, it is possible to use i.e. MySQL statements to get automated analysis using R-scripts.
The second aspect we report is about design of an experiment involving large data collection procedures: in our case for around 76% from the originally 1775 selected cows we got all desired information at the end. It depends on the duration and kind of an experiment how much more individuals must be selected to get the desired complete dataset of an experiment.
Our third contribution is to deliver an example of how to design a GC-MS metabolite profiling taking into account a number of influencing factors for a complex study like our example.

References

James, D. A. and DebRoy, S. (2004). R interface to the MySQL database. *http://cran.r-project.org/doc/packages/RMySQL.pdf*.

Lisec, J., Schauer, N., Kopka, J. *et al.* (2006). *Nat. Protoc.*, 1:387–396.

Meuwissen, T. H., Hayes, B. J., and Goddart, M. E. (2001). *Genetics*, 157:1819–29.

R Development Core Team (2005). R: A language and enviroment for statistical computing. *http://www.R-project.org*.