CHAPTER 6

Conclusions and Perspectives

In Chapter 2 of this thesis, similarities and differences among members of (mainly MZ) twin families in their blood plasma lipidomics profiles were investigated. The results of these analyses suggested that shared genetic background and shared environmental experiences contribute to similarities in blood plasma lipidomics profiles among individuals. Male and female participants segregated almost perfectly at the highest level in the dendrogram resulting from hierarchical clustering analysis. Clustering of MZ co-twins was assessed by counting the number of branching points in the dendrogram separating both twins, and comparing the observations with reference distributions based on permutation testing. Indeed, based on these comparisons it could be concluded that in general more MZ twins belonging to the same twin pair clustered together than was expected on the basis of chance. However, for some MZ twin pairs the distances between co-twins were larger than was expected on the basis of their genetic similarity. Such dissimilarity of lipid profiles between MZ co-twins appeared to correlate positively with female gender, relatively high CRP concentration and, in a number of cases, with recent illness.

In Chapter 3, a data transformation method was presented to make combinable (with the variables as the shared mode) data sets obtained with the same semiquantitative analytical chemical method but in different measurement "blocks". Such "blocks" can arise, for example, when the measurements of all samples for a particular study can not be performed at the same time. The application of the data transformation method, referred to as "quantile equating", was demonstrated with data sets obtained by LC–MS analysis of blood plasma lipids, and by ^{1}H NMR spectroscopy of blood plasma and urine samples from twin families.

The combined LC–MS data sets obtained after application of the "quantile equating" method described in Chapter 3, were used for the analyses described in Chapter 4. In this Chapter it was demonstrated in hierarchical clustering analysis that quantile equating had indeed been beneficial for making the LC–MS data sets combinable. Furthermore, on the basis of this larger data set including notably more DZ twin families, the general findings described in Chapter 2 could be replicated. That is, the results described in Chapter 4 also supported the hypothesis that shared genetic background and shared environmental exposure contribute to similarities in lipidomics profiles among individuals. Also, in general dissimilarities in lipidomics profiles between female MZ co-twins were larger than between male MZ co-twins. However, the positive correlation between dissimilarity of lipid profiles between MZ co-twins, recent illness and relatively high CRP concentration was not as apparent as on the basis of the analyses described in Chapter 2.

Finally, Chapter 5 describes the results of uni- and multivariate quantitative genetic analyses of blood plasma LC–MS and ^{1}H NMR data on the basis of structural equation modeling. Univariate analyses of the LC–MS data, which were generated using a "targeted" method for the analysis of lipids, suggested different patterns of heritability for lipids belonging to different lipid classes. Interestingly, within the triglyceride class we observed different heritabilities for lipids with different numbers of C-atoms and/or different numbers of double bonds in the fatty acid backbone. The dendrogram resulting from hierarchical clustering analysis of the genetic correlations among all lipids suggested shared genetic factors contributing to the phenotypic covariance of lipids from the same lipid class. The heritabilities of the features detected in the ^{1}H NMR data, which were generated using a "global" method to obtain an overview of metabolites from different classes, displayed much larger heterogeneity with respect to those of the lipids detected with LC–MS. Also, considerable heterogeneity was observed in the genetic correlations among all features, which was again as expected on the basis of the "global" nature of NMR spectroscopy.

6.1 Between-block effect correction methods in metabolomics

The method described in Chapter 3 of this thesis appears to be one of the first to address the issue of "between-block" effect correction with application to semi-quantitative analytical chemical data. It is argued in Chapter 3 that systematic nonbiological differences between semi-quantitative data obtained in different measurement "blocks" can exist, for example due to small analytical changes between the blocks that are not avoidable by good analytical practice alone. The method of univariate "quantile equating" is introduced to address this issue when there are nonlinear differences between the distributions of the data obtained on the same variables in different measurement blocks.

That "between-block" effect correction at "low" level (*i.e.*, at data level)

appears to be a relatively unexplored area of research in the context of semiquantitative metabolomics measurements, is somewhat surprising in view of the large number of publications on similar topics within the transcriptomics field. In transcriptomics, the analogue of what we in Chapter 3 of this thesis refer to as "between-block" effects is often referred to as "batch effects". Several authors[186–192] give similar considerations to correct for "batch effects" in microarray studies, as we do for correcting for what is called "between-block effects" in Chapter 3. Demetrashvili *et al.*[186] applied the empirical Bayes method of Johnson *et al.*[191] to correct for "batch effects" after application of the loess normalization within arrays, which implies that normalization alone was not sufficient in their case for between-batch effect correction. Other authors have described similar findings.[189,191] This reported insufficiency of normalization to correct for between-batch effects in microarray studies is in concordance with our finding that it is not sufficient for correction for between-block effects in metabolomics data. Jiang and colleagues[189] developed the "disTran" method for between-batch effect correction of microarrays, which is probably equivalent to our "quantile equating" method that we used for between-block effect correction in the context of a metabolomics study. Several authors (*e.g.*,[193]) have even presented methods to make combinable (with the variables as the shared mode) data sets obtained with different gene expression measurement techniques.

The difference in nomenclature employed in the context of microarray studies (*i.e.*, "batch effect correction") and in the context of semiquantitative metabolomics studies (*i.e.*, "between-block effect correction") might reflect a difference in application domain of highly similar data pretreatment methods. Indeed, the severity of "batch effects" as generally described within the context of metabolomics studies, appears to be relatively limited with respect to that of the "batch effects" described for microarray studies. Therefore, in metabolomics studies, data obtained in different batches but within the same "block" are often reported to be combinable either without correction, or with batch effect correction using for example repeatedly measured quality control samples.[2,117,128,194] However, apparently in contrast to the situation within gene expression studies, the possibility and even necessity to consider data pretreatment techniques for between-block effect correction does not appear to be accepted yet by the metabolomics community. Rather, currently there seems to be a preference for perfection of the stability and robustness of the used analytical chemical platforms, such that data obtained with the same analytical chemical method in different measurement blocks can be combined without additional correction. For example, efforts are being undertaken to standardize working protocols.[2,21,195–197] However, among transcriptomics researchers a keen interest in methods that correct for "batch effects" still exists, despite similar efforts in that field.[187] With the demand to discover biological effects of ever smaller effect size on the basis of metabolomics data,[117] it is foreseeable that the application domain of methods to correct for "between-block" effects increases in response to this demand as well.[198]

Finally, a caveat for the application of methods for block effect correction to semiquantitative metabolomics data sets might be in place. Currently complete identification of all detected compounds in metabolomics studies is often not possible.[21] The LC–MS data discussed in this thesis, for example, were based on an analytical method that cannot distinguish among different isomers of a detected lipid.[127] Therefore, it could not be verified whether for example the ratios of different isomers of the 'same' lipid in data sets originating from different measurement blocks were equal. However, an important assumption when applying "equating" methods to make combinable data sets, is that data from the same variables (*e.g.*, the same isomers of a particular lipid) are equated in different data sets. Any indications that this assumption might be violated in a given study might preclude the application of equating methods in order to avoid bias. Nevertheless, it is concluded that useful methodology to correct for batch and/or block effects in semi-quantitative metabolomics studies might be adopted from microarray research. A similar case was made by Redestig *et al.*[199]

6.2 Multivariate quantitative genetic analysis

In Chapters 2 and 4 of this thesis, multivariate quantitative genetic analysis was performed based on the distances among objects, computed on the basis of blood plasma lipidomics profiles. In Chapter 5, multivariate quantitative genetic analysis was performed on the basis of structural equation modeling. In Chapters 2 and 4, we have used the 'unsupervised', hypothesis-free data analysis method of hierarchical clustering. As has been explained in the General Introduction, the aim of hierarchical clustering analysis is to "see what the data are trying to tell us".[41] Nevertheless, the results in Chapters 2 and 4 were consistent with our hypothesis that shared genetic background and shared environment contribute to similarities in blood plasma lipidomics profiles among individuals.

Structural equation modeling, which was used in Chapter 5, is initiated by the specification of a model that formalizes a hypothesis about the causal relationship between predictors and predicted variables. Hence, structural equation modeling could be regarded a 'hypothesis-driven' method. However, in Chapter 5 we have used structural equation modeling in a relatively hypothesis-free way. That is, a structural model based on Cholesky composition of the variance component matrices was used, which is a relatively hypothesis-free model.[29] Also, the genetic correlations for all pairs of variables, estimated using this hypothesis-free model, were analyzed using the 'unsupervised', hypothesis-free method of hierarchical clustering. Nevertheless, the patterns of clustering of lipids on the basis of their genetic correlations were consistent with the hypothesis that metabolites from the same metabolite class correlate positively because of shared genetic factors of phenotypic variation.

This methodology for multivariate quantitative genetic analysis on the basis

of SEM might be further enhanced by the development or application of methods that allow the joint analysis of all variables in one multivariate analysis, rather than the 'multistep multivariate' approach. That is, from a purely mathematical point of view, the results from "multiple bivariate" analyses cannot be jointly analyzed because they are not in the same multivariate space.

Furthermore, as explained below, the results of the analyses based on the distances among objects could provide indications which 'moderators' might be placed where in a structural equation model to be used for quantitative genetic analysis. In structural equation modeling, moderators are covariates that influence for example the weight of predictor variables.[200] It can be hypothesized, for example, that gender 'moderates' the relative contribution of genetic variance to phenotypic variance and such a hypothesis can be formalized as a moderator on the path coefficients in a structural equation model. The analyses based on the distances among objects, as in Chapters 2 and 4 of this thesis, might be used to explore the heterogeneity among the individuals in the study sample, to find indications whether there are potential covariates that might be included as moderators in a structural equation model. For example, in Chapter 2 in hierarchical clustering analysis we observed almost perfect segregation of male and female participants at the highest level in the dendrogram. This suggests that gender might be included as a covariate on the means in structural equation models.

6.3 Medical relevance of our findings

In Chapters 2 and 4, individual differences were studied on the basis of distances among objects (lipidomics profiles) in multivariate space. The results of these analyses suggested that for example disease might increase such individual differences in blood plasma lipid concentrations. Indeed, our results on the basis of blood plasma lipid profiling support the hypothesis that "because of biological individuality, each individual will have a particular location within the larger distribution of quantitative values that describe the parameter in the population; the private homeostatic value may then be seen to be displaced because the individual's system is [...] overwhelmed by experience".[201] The power to detect the effects on individual differences of particular important factors, such as disease, might be enhanced in analyses on the basis of distances among objects with respect to univariate analysis. This increase in statistical power should be due to the fact that in the multivariate distances among objects, the effects of factors that influence phenotypic variation in the individual variables (as can be assessed for example in univariate analysis on the basis of structural equation modeling, as was performed in Chapter 5) are "pooled". This "pooling" occurs during the summation of the dissimilarities among the objects for the individual variables (see for example equation 1.5 in the General Introduction). Further studies are necessary to determine the magnitude of this gain in statistical power due to studying distances among

objects rather than studying the variation in individual variables.

The results of the univariate analyses based on structural equation modeling as described in Chapter 5 of this thesis are informative of the relative contribution of genetic variation and environmental variation to the quantitative variation in individual metabolites.

The genetic correlations as estimated in the multivariate quantitative genetic analyses described in Chapter 5 are informative of the genetic structure that underlies the phenotypically observable quantitative relationships among different metabolites. These results might be relevant for the study of common diseases,[47,66] and might enhance the interpretation of the findings from *e.g.* GWA studies.