Analysis of Metabolomics Data from Twin Families

Hermanus H.M. (Harmen) Draisma

Analysis of metabolomics data from twin families
Hermanus H.M. (Harmen) Draisma
PhD thesis with summary in Dutch
ISBN: 978-90-745-3875-6
Typeset in LaTeX 2_ε
Produced by F&N Boekservice

Cover: Constellation of Gemini (twins) represented as a metabolite in metabolite space. Image from the stellar atlas *Firmamentum Sobiescianum sive Uranographia* (Johannes Hevelius, 1690) kind courtesy of the U.S. Naval Observatory and the Space Telescope Science Institute.

Analysis of Metabolomics Data from Twin Families

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 10 mei 2011
klokke 16:15 uur

door

Hermanus H.M. (Harmen) Draisma
geboren te Voorburg
in 1981

Promotiecommissie

Promotores:

Prof.dr. J. van der Greef
Prof.dr. T. Hankemeier
Prof.dr. J.J. Meulman

Copromotor:

Dr. T.H. Reijmers

Overige leden:

Prof.dr. D.I. Boomsma (Vrije Universiteit Amsterdam)
Prof.dr. M. Danhof (Universiteit Leiden)
Prof.dr. G.J.B. van Ommen (Leids Universitair Medisch Centrum)
Prof.dr. P.E. Slagboom (Leids Universitair Medisch Centrum)
Prof.dr. A.P. IJzerman (Universiteit Leiden)

The research described in this thesis was performed at the Division of Analytical Biosciences of the Leiden/Amsterdam Center for Drug Research, Leiden University, Leiden, The Netherlands. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Printing of this thesis was supported financially by the Leiden/Amsterdam Center for Drug Research, and by the Netherlands Bioinformatics Centre.

Whoever saves a life, it is considered as if he saved an entire world

Mishnah Sanhedrin 4:5; *Babylonian Talmud* Tractate Sanhedrin 37a

Contents

CHAPTER 1

General Introduction

1.1 Something old, something new: twin studies and metabolomics

The elucidation of the relative importance of genes and environment for variation in traits using data from twins and families has a long history.[1] Metabolomics, or the study of small molecules that are the reactants, intermediates or end products of cellular metabolism, on the other hand, is a relatively young field within the "omics" sciences.[2] This thesis describes the results of various analyses that address different questions that may arise within the context of analysis of metabolomics data from twin families.

In this General Introduction, first the concepts of metabolomics, and of twin and family studies are introduced. Then, two approaches are discussed that can be used for the analysis of multivariate data, such as metabolomics data, as obtained from families. These two approaches, *i.e.* structural equation modeling and hierarchical clustering analysis, are central in this thesis because they can both be informative of the contributions of genetic and environmental variation to variation in metabolite levels. Finally the value of our approach in the context of the recent developments in genome-wide association studies is discussed. A short outline of the remainder of this thesis is given at the end of this Introduction.

1.2 Metabolomics

"Omnia mutantur nihil interit" — everything changes but nothing is truly lost. Reportedly, even more classic than these legendary words attributed to Pythagoras by Ovid in his *Metamorphoses* (AD 8)[3] is the notion that metabolites can be informative of the status of organisms. For example, approximately two millennia before Ovid completed his masterwork, Chinese doctors used ants to detect high glucose levels in urine as an indicator for diabetes.[4,5]

Whereas Ovid's book describes a number of cases where changes in form or shape had been effectuated by witchcraft, *i.e.* metamorphoses, the term "metabolism" also refers to a change but to one which can be studied by scientific means. The "changes" that constitute metabolism are the conversions of typically low-molecular weight molecules into other molecules due to the actions of enzymes and in some cases also co-factors. The molecules that are the substrates, intermediate or end products of metabolism are called metabolites.[2] One could argue that rather than being lost, the study of metabolism is on the contrary of increasing interest because it is conceived that metabolic processes are particularly directly linked to the functioning of cells, organs, and even complete organisms.

The central dogma in molecular biology dictates that information flow within cells goes from genes, via gene transcripts, to proteins.[6,7] In this view, genes encode the heritable information that is transmitted from parent to child; this information is transcribed from DNA into messenger RNA, which in turn encodes the sequence of amino acids in proteins. Enzymes are a subclass of proteins, some of which can convert metabolites into other metabolites. The metabolome (*i.e.*, the complement of all metabolites) has been recognized by several authors[8,9] to be an integral part of the molecular biological central dogma as well. Furthermore, it is increasingly recognized that there is considerable cross-talk between the different information levels in the central dogma, and that therefore the view of unidirectional flow of information as proposed by the central dogma is probably too simplistic.[2,9] This leads to a view of different information levels and their interrelationships within cells as depicted in Figure 1.1.

Despite the crosstalk among the different physiological levels at which cellular functioning can be studied, the metabolome is conceived to be the level that is relatively the closest to the outward measureable characteristics, such as physiological functioning, of the cell. Such outward measureable characteristics are referred to as "phenotypes", and because metabolites are physiologically in between the genome and the phenotypic appearance of, for example, a cell, metabolites are sometimes called "endophenotypes" or "intermediate phenotypes".[5,10–12] Figure 1.1 also shows that next to genetic variation, the influence of the environment is important for the resulting phenotype at all information levels.

Whereas the recognition of metabolites as being informative of the state of biological entities has certainly not been lost, what indeed has changed is the

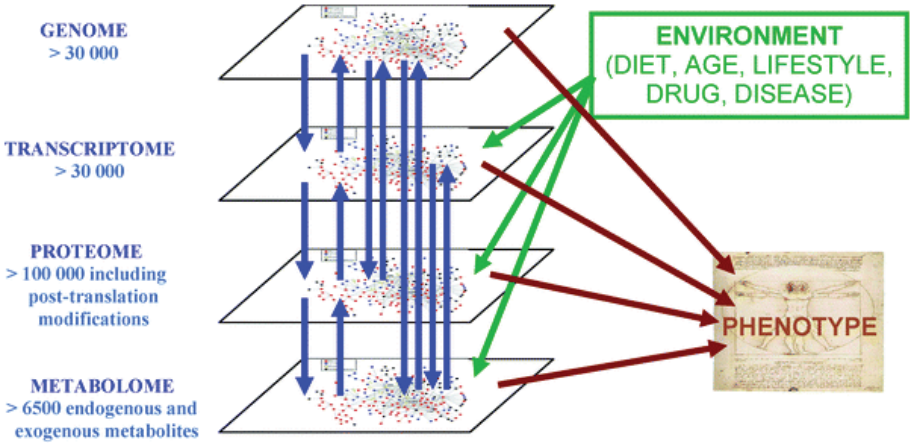

Figure 1.1: The different information levels within (human) cells and their interactions. The phenotype is a function of the interactions of the components of genome, transcriptome, proteome, and metabolome with each other and with the environment. According to this view, the environment might exert modulating effects on all levels except the genome. Reproduced from Dunn *et al.*[2] by permission of The Royal Society of Chemistry.

way we are able to look at metabolism and metabolites. Science has come a long way since the earliest clinical applications of metabolite measurements. Due to significant technological advances mainly in the second half of the 20$^{\text{th}}$ century, it is now possible to comprehensively measure large numbers of different metabolites in a given sample. Studies that employ such comprehensive, or holistic, measurement approaches are often referred to as "metabolic profiling", "metabolomics", "metabolite fingerprinting", or "metabonomics".[2,9] Although most definitions acknowledge differences in *e.g.* the application domain, numbers of detected metabolites, and technology among the disciplines indicated with the different terms, the designation "metabolomics" is probably the most widely used. The suffix "omics", in analogy with for example "genomics" (study of genetic variation) and "transcriptomics" (study of gene expression), indicates the comprehensive nature of the approach.[13,14] The work described in this thesis can be described as human metabolomics studies; however, notably, also microbial,[15] plant[16] and animal[17] metabolomics exist. The remainder of this discussion of metabolomics will focus on human metabolomics.

Metabolomics studies should follow the steps of a general workflow.[5,18–20] Ideally, a metabolomics study starts with a biological question. This does not mean, however, that there is always a clear hypothesis about the biological effects that will be observed. For example, based on existing knowledge about a particular disease, it might be suspected that lipids play an important role; then this could warrant the choice for a targeted lipidomics platform for measurement of samples from study participants with and without the dis-

ease. However, it may not be hypothesized *a priori* which specific biological pathways might be involved in causing the disease. Rather, on the basis of the variation observed in the data obtained by metabolomics measurements, such specific hypotheses with respect to the involvement of particular pathways might be generated. Indeed, as such metabolomics is typically a "data driven" and "hypothesis-generating" discipline.

On the basis of the biological question, an experimental design is formulated. For a typical human metabolomics study, often samples from "biofluids" such as blood or urine but sometimes also *e.g.* cerebrospinal fluid or saliva are obtained from healthy volunteers and/or from patients. The samples are first processed, for example to extract the most relevant classes of metabolites for that particular study. This is a way to 'target' the analysis at these particular classes, for example only at the lipids in a blood sample. On the other hand, in a 'global' analysis the aim is to obtain an overview of the (relative) concentrations of metabolites from all classes present in a given sample, such as amino acids, lipoproteins, and carbohydrates.

For detection of metabolites, nowadays ants have been replaced by for example mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. A mass spectrometer detects the ratio of mass versus charge of an ionized metabolite.[2] In a frequently used type of NMR spectroscopy, *i.e.* proton NMR spectroscopy (^{1}H NMR), simply stated the energy is detected as it is emitted by metabolites depending on the chemical environment surrounding the protons in the molecule. A 'global' analysis of a blood plasma sample using NMR spectroscopy, for example, will typically require less sample processing than a 'targeted' lipidomics analysis using liquid chromatography–mass spectrometry (LC–MS). In both global and targeted analyses, the different metabolites present in the preprocessed sample can be separated before detection to enhance the ability to detect them. For example, a chromatographic separation step such as gas chromatography (GC) or liquid chromatography (LC) can be used prior to detection in order to separate the different metabolites based on their differences in physicochemical characteristics (*e.g.*, hydrophobicity).

The advances in the development of analytical techniques allow for the measurement of hundreds to thousands of different metabolites in a given sample. The data resulting from such measurements typically require additional preprocessing before they can be subjected to statistical analysis. Such preprocessing can be for example the extraction of peaks corresponding to different metabolites from an NMR spectrum. The heights or integrals of these peaks can then be collected into data tables where for each measured sample the heights or integrals of all peaks are listed.

Ideally, one would like to obtain quantitative measurements (concentrations) of all the metabolites in a given sample; however, amongst others due to the complexity of most samples this is often not possible.[2,21] Therefore many metabolomics studies are semiquantitative rather than quantitative, implying that the relative concentrations of different metabolites with respect to each other can be measured, but not the absolute concentrations. Another current

challenge in metabolomics is identification: often the identity (structure) of the detected compounds cannot be resolved completely.

Often the data have to be pretreated to make them comparable across different samples using uni- or multivariate statistical techniques. An example of such data pretreatment is sample-wise normalization; for instance, the sum of the integrals of all peaks for each sample can be made equal to one, to acknowledge that differences among samples with respect to this sum are not biologically relevant.[22]

Subsequently, the data in these tables can be subjected to statistical analysis. Because a typical metabolomics study aims to comprehensively detect either a large number of metabolites from different classes, or all metabolites of a given class, there will often be multiple different metabolites that display similar changes due to a particular biological effect. For example, when comparing the metabolite profiles of healthy volunteers with those from study participants with a particular disease, the aim of statistical analysis of the metabolomics data might be to find patterns of metabolites that display similar differences in concentration between healthy and diseased individuals. Such metabolites that indicate a change from healthy to diseased are often called "biomarkers".

Because it is often expected that in metabolomics studies a biological effect of interest will manifest as changes in multiple related metabolites, for statistical analysis in particular multivariate techniques are used. A multivariate statistical technique typically acknowledges that a particular biological effect can manifest as the linear combination of the effects observed in multiple different individual metabolites. For example, principal component analysis (PCA)[23] is a multivariate statistical technique that can be used to uncover the direction of the dominant variation exhibited by multiple individual metabolites. An important advantage of the use of multivariate statistical techniques in metabolomics research is that these techniques, by taking into account the information present in all variables rather than in one variable at a time, are statistically much more powerful than univariate statistical methods. Therefore, using multivariate techniques statistical inference is often possible in metabolomics on the basis of much smaller numbers of measured samples than would be the case with univariate analysis, provided that the results are sufficiently validated.[24]

1.3 Twin and family studies

The origin of family studies to elucidate the relative influences of genetic variation and environmental variation on phenotypic variation dates back to Sir Francis Galton (1822–1911). In his 1869 book "Hereditary Genius — An Inquiry into its Laws and Consequences",[25] Galton describes his finding that, starting with "illustrious men" as probands, close relatives of such men displayed remarkable genius as well, but that these phenotypic similarities de-

creased when comparing more distant relatives. Galton also recognized that statistical analysis of such data obtained in family members was key to derive what he referred to as "a decided law of distribution of genius in families".[25]

Although the historical importance of Galton's initial findings can hardly be disputed, these findings merely showed that characteristics 'run in families', *i.e.* that family members will be more similar than non-family members for a given phenotype such as intelligence. However, Galton's publication five years later of another book[26] illustrates that indeed he had become aware of the distinction between "nature and nurture", or, in other words, of the distinction between genetic and environmental influences on phenotypic values.

The pioneer work of Galton in general families was later expanded and its potential was enhanced by acknowledging that in particular using twin families, the power to detect genetic effects is large. Therefore, studies of twins and families have traditionally been very important within the field of quantitative genetics.[27]

Quantitative genetics is the study of the genetic causes of individual differences for measurable traits.[28] Examples of such traits are height, weight, depression, migraine, or metabolite levels as measured in body fluids of normal humans. In general, in quantitative genetics the following genetic and environmental sources of phenotypic variance are considered: additive genetic effects ("A"), non-additive genetic effects ("D", for 'dominance'), common or shared environmental effects ("C"), and specific non-shared environmental effects ("E"). The term "additive genetic effects" is used to refer to genetic effects where the total effect equals the sum of the effects of alleles that influence the value for a trait.[29] "Non-additive" genetic effects are not simply the sum of the effects of alleles, because of interactions within or between loci. Well-known examples of non-additive genetic effects are dominance (within locus) and epistasis (across loci). Common or shared environmental effects are the environmental effects shared by members of the same family; an example is diet.[30] Specific environmental effects are not shared by relatives; measurement error is also included in this source of phenotypic variance.[29]

A classic method within the field of quantitative genetics is the "classical twin study", which relies on the comparison of phenotypic similarities between monozygotic (MZ) and dizygotic (DZ) co-twins raised together for estimating the relative importance of these respective sources of phenotypic variation. MZ co-twins share 100% of their additive genetic variation ("$varA$"), whereas this percentage is on average 50% between DZ co-twins; this latter percentage is the same for biological nontwin siblings. In twin studies, it is assumed that MZ and DZ co-twins share the same degree of common environmental variation ("$varC$").[27] The basic idea behind the classical twin study is that therefore, any excess phenotypic correlation between MZ co-twins over that between DZ co-twins must be due to genetic effects.[1] More formally, the difference in the degrees of shared genetic variation between MZ co-twins and between DZ co-twins can be used in statistical analysis to disentangle the genetic variation from the environmental variation influencing the individual differences in trait

values measured in these twin pairs.

The combination of a large difference in shared additive genetic effects between MZ and DZ co-twins, and the same degree of shared environmental variation in MZ and DZ co-twins, causes the classical twin study to have an importantly increased statistical power over that of nontwin family-based quantitative genetic analyses to detect genetic components of phenotypic variance. [31]

1.4 Two alternative methods to separate "nature" from "nurture" using family data

Quantitative genetic analysis using structural equation modeling (SEM) is a classic method to estimate genetic variation and environmental variation for phenotypes observed in family data. However, in this thesis an alternative method is described for quantitative genetic analysis based on hierarchical clustering analysis of multivariate family data. Although both analysis of hierarchical clustering among family members, and multivariate quantitative genetic analysis by SEM can be considered to be multivariate statistical techniques, they aim to answer different questions with respect to "nature" (genes) versus "nurture" (environment). Below, a general introduction is given into SEM and hierarchical clustering analysis in the context of the quantitative genetic analyses described in this thesis.

1.4.1 Structural equation modeling

SEM is a generic statistical technique for testing hypotheses. [27,32,33] As Bollen (1989) states: "The purpose [of SEM] is to determine if the causal inferences of a researcher are consistent with the data". [32]

The initial step in SEM is the generation of a "structural model" that formalizes a hypothesis of the causal relationship between variation in predictor variables and variation in predicted variables. This "structural model" can be represented as a set of "structural equations", or equivalently as a so-called "path diagram". Path analysis was originally developed for genetic analysis by Sewall Wright. [34]

The fit of this model to observed data is optimized by iteratively changing the values of the free parameters in the structural model. Often maximum likelihood is used to obtain parameter estimates. [35] It is customary to compare the likelihoods of different versions of the initial model that vary in complexity, with the aim to identify the model that yields the best trade-off between model complexity and fit to the observed data. With likelihood ratio tests one can test whether the likelihood changes significantly when fitting a nested model.

In this thesis, SEM is used to obtain estimates for the relative contribution of genetic and environmental factors to the phenotypic variation as observed in twin families. Because in such analyses the structural equations often formalize a hypothesis of the causes of the covariance observed in the measured data, this

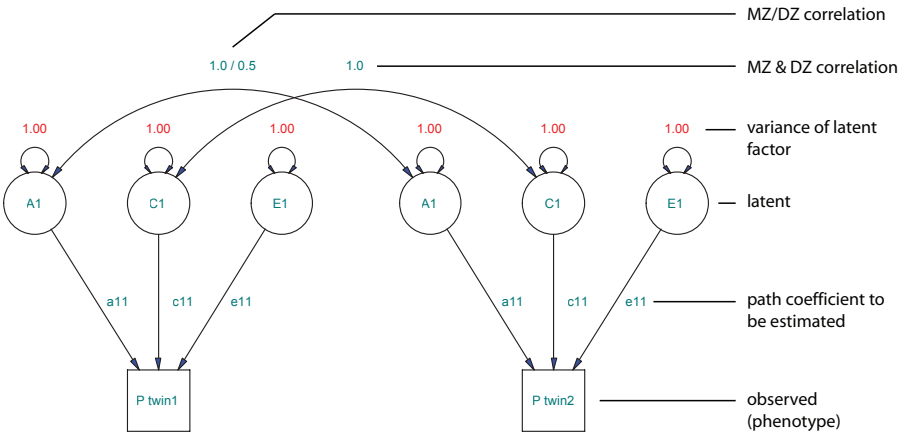

Figure 1.2: Path diagram for univariate quantitative genetic analysis under classical twin design. Latent (unobserved) factors and path coefficients are indicated by uppercase (*e.g.*, "A") and lowercase (*e.g.*, "a") letters, respectively. In a path coefficients model, the variances of the factors are standardized to a value of one,[27] as indicated in this figure. Note the different coefficient values for the additive genetic covariance components in MZ and DZ co-twins of 1.0 and 0.5, respectively. A1, C1 and E1, latent additive genetic, common environmental and specific environmental factors, respectively. "P twin1" and "P twin2", phenotype (trait) in first and second members of twin pairs, respectively.

type of SEM is also referred to as "analysis of covariance structures".[32]

An initial question in twin and family studies is often whether genetic ("*A*", "*D*") or environmental ("*C*", "*E*") influences are more important for the variation observed in a single trait. The aim of univariate quantitative genetic analysis on the basis of SEM is to answer this question. Of note, on the basis of the classical twin design (pairs of MZ and DZ twins reared together), the separate contributions of "*C*" and "*D*" cannot be estimated on the basis of a model that includes both sources of variance.[29]

An example of a structural model for univariate analysis under the classical twin design is given in Figures 1.2 and 1.3. As explained below, Figure 1.2 can be considered the graphical representation of the equivalent covariance structure as depicted in Figure 1.3. The path diagram as depicted in Figure 1.2 can be conceived as to represent a model of the relationship between unmeasured latent factors ($A1$, $C1$, $E1$) and the phenotype (P) as measured in a single individual. For example, if we assume that the phenotypic data for all measured individuals have been reduced to "mean deviation form" (*i.e.*, the mean phenotypic value has been subtracted from the values of all individuals),[27] the phenotypic score P_i for an individual i (*i.e.*, the deviation of his or her phenotypic value from the mean phenotypic value over all individuals) can be described to be a function of this individual's genetic and environmental factor

$$\Sigma_{MZ} = \begin{bmatrix} a_{11}^2\text{varA} + c_{11}^2\text{varC} + e_{11}^2\text{varE} & a_{11}^2\text{varA} + c_{11}^2\text{varC} \\ a_{11}^2\text{varA} + c_{11}^2\text{varC} & a_{11}^2\text{varA} + c_{11}^2\text{varC} + e_{11}^2\text{varE} \end{bmatrix}$$

$=$

	P twin1	P twin2
P twin1	$a_{11}^2 + c_{11}^2 + e_{11}^2$	$a_{11}^2 + c_{11}^2$
P twin2	$a_{11}^2 + c_{11}^2$	$a_{11}^2 + c_{11}^2 + e_{11}^2$

$$\Sigma_{DZ} = \begin{bmatrix} a_{11}^2\text{varA} + c_{11}^2\text{varC} + e_{11}^2\text{varE} & 0.5 \times a_{11}^2\text{varA} + c_{11}^2\text{varC} \\ 0.5 \times a_{11}^2\text{varA} + c_{11}^2\text{varC} & a_{11}^2\text{varA} + c_{11}^2\text{varC} + e_{11}^2\text{varE} \end{bmatrix}$$

$=$

	P twin1	P twin2
P twin1	$a_{11}^2 + c_{11}^2 + e_{11}^2$	$0.5 \times a_{11}^2 + c_{11}^2$
P twin2	$0.5 \times a_{11}^2 + c_{11}^2$	$a_{11}^2 + c_{11}^2 + e_{11}^2$

Figure 1.3: Covariance structures for univariate quantitative genetic analysis under classical twin design. "*varA*", "*varC*", "*varE*", additive genetic, common environmental, and specific environmental variance components, respectively. "*a*", "*c*", "*e*", path coefficients. Path coefficients are computationally equivalent to standard deviations; hence their squared values constitute the variance components as is evident from the covariance structures as depicted schematically in this figure. Note the different coefficient values for the additive genetic covariance components in MZ and DZ co-twins of 1.0 and 0.5, respectively. "P twin1" and "P twin2", phenotype (trait) in first and second members of twin pairs, respectively; Σ_{MZ} and Σ_{DZ}, expected covariance matrices for MZ and DZ co-twins, respectively. A similar coloring as in Figure 4 of[29] is used here, *i.e.*, cells representing within- and cross-twin (co)variances are colored light and dark grey, respectively.

scores and of the factor loadings as follows:

$$P_i = a_{11}A_i + c_{11}C_i + e_{11}E_i \tag{1.1}$$

where P_i is the phenotypic score, a_{11} is the loading of the additive genetic latent factor, A_i is the score of this individual on the additive genetic latent factor, and analogously for the common and specific environmental factor loadings (c_{11} and e_{11}) and scores (C_i and E_i). The values of the factor loadings a_{11}, c_{11} and e_{11} are the same for all individuals i. Hence, these factor loadings can be regarded to be the weights, or regression coefficients, that are assigned to the scores on the latent factors in order to produce the phenotypic scores.

Two types of model specification are possible: one in which the values of the factor loadings are fixed to have the same value for all latent factors, and the variances of the latent factors are allowed to vary freely; and the other type of specification in which the variances of the latent factors are fixed to have the same standardized value of 1 and the values of the path coefficients are allowed to vary freely. The model depicted in Figure 1.2 represents the latter situation, known as the "path coefficients model".[27] In this case, the variance of the phenotypic scores over all individuals can be represented as follows:[32]

$$\begin{aligned} varP &= a_{11}^2 \cdot varA + c_{11}^2 \cdot varC + e_{11}^2 \cdot varE \\ &= a_{11}^2 \cdot 1 + c_{11}^2 \cdot 1 + e_{11}^2 \cdot 1 \\ &= a_{11}^2 + c_{11}^2 + e_{11}^2 \end{aligned} \tag{1.2}$$

In general, the elements on the diagonal of a covariance matrix are variances, and the off-diagonal elements are covariances.[36,a] The entries on the diagonals of the expected covariance matrices as in Figure 1.3 represent the variances of a vector of values observed for the trait in the first (upper left) and second (lower right) members of a number of twin pairs. The off-diagonal elements in the expected covariance matrices in Figure 1.3 represent the covariances between the vectors of values observed for the trait in the first and the second members of twin pairs.

The proportions of the respective sources of phenotypic variance shared by individuals are specified by coefficients in the structural equation model. For example, the additive genetic correlation between MZ co-twins is 1.0, because MZ co-twins share (nearly) all genetic variance at the DNA sequence level. For DZ co-twins, this correlation is equal to 0.5; hence the coefficient values of 1.0 and 0.5 for "A1" in the elements of the expected covariance structure representing the covariance between MZ co-twins and between DZ co-twins, respectively.

The fit of the expected covariance matrix (computed on the basis of the model) to the 'observed' covariance matrix (*i.e.*, the covariance matrix computed from the observed data) is optimized by iteratively changing the values

[a]Note that a covariance matrix is equivalent to an unstandardized correlation matrix; hence it can easily be seen that covariances (the off-diagonal elements of a covariance matrix) will often have lower values than variances (the diagonal elements of a covariance matrix)

of the parameters "a_{11}", "c_{11}" and "e_{11}" that form the basis for the expected covariance matrix. The parameter values that yield the best fit of the expected covariance matrix to the observed covariance matrix are considered to be the estimates under the specified model for the additive genetic, common environmental and specific environmental effects that constitute the phenotypic variance observed in the studied population sample.

The standardized (with respect to total phenotypic variance) genetic variance component as resulting from a univariate analysis is often called "heritability". If the distinction is made between additive and non-additive genetic effects, then the proportion of phenotypic variance attributable to additive genetic effects is called "narrow heritability". If this distinction is not made, then the proportion of phenotypic variance attributable to all genetic effects together is called "broad heritability".[37]

Next to a univariate analysis as described above, which gives estimates for the relative influences of genetic and environmental variation for the variation observed in a given trait, multivariate quantitative genetic analysis can be used to elucidate the relative importances of shared (among traits) genetic and shared (among traits) environmental variance for the observed *covariance* among two or more traits. Figures 1.4 and 1.5 depict the path diagram and the corresponding covariance structure for a relatively simple, hypothesis-free[29] bivariate analysis based on the classical twin design.

In the model depicted in Figure 1.5, the variance component matrices "$varA$", "$varC$", and "$varE$" that constitute the expected covariance matrix are computed as the products of lower triangular matrices of path coefficients and their transpose. Let us take the submatrix of the expected covariance matrix representing the covariance between DZ co-twins as an example; we will denote this submatrix as Σ_{DZsub}. Analogous to the covariance between DZ co-twins in the univariate example (see Fig. 1.3), this submatrix Σ_{DZsub} of the expected covariance matrix is computed as the algebraic sum of the variance component matrices "$varA$" (multiplied by the DZ covariance coefficient value of 0.5) and "$varC$" (cf. Fig. 1.5):

$$\Sigma_{DZsub} = 0.5 \times varA + varC \tag{1.3}$$

$=$

Twin2 \ Twin1	P1	P2
P1	$0.5 \times a_{11}^2 + c_{11}^2$	$0.5 \times a_{11} \times a_{21} + c_{11} \times c_{21}$
P2	$0.5 \times a_{11} \times a_{21} + c_{11} \times c_{21}$	$0.5 \times a_{22}^2 + 0.5 \times a_{21}^2 + c_{22}^2 + c_{21}^2$

Taking "$varA$" as an example, the matrices representing the expected variance

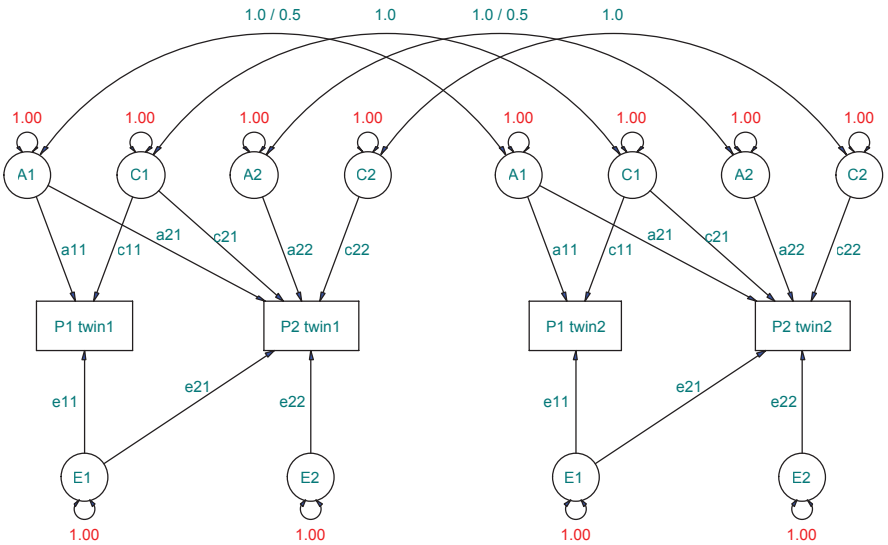

Figure 1.4: Path diagram for bivariate quantitative genetic analysis under classical twin design. A1 and A2 are additive genetic factors, where A1 is possibly shared by both traits and A2 is specific for phenotype 2, respectively; similarly for C1 and C2 (familial environment), and for E1 and E2 (specific environment). Note again the different correlations for the additive genetic factors in MZ and DZ co-twins of 1.0 and 0.5, respectively. P1 and P2, phenotype (trait) 1 and 2, respectively; twin1 and twin2, first and second members of twin pairs, respectively.

$$\Sigma_{MZ/DZ} = \begin{bmatrix} a_{11}^2\text{varA} + c_{11}^2\text{varC} + e_{11}^2\text{varE} & 1.0/0.5 \times a_{11}^2\text{varA} + c_{11}^2\text{varC} \\ 1.0/0.5 \times a_{11}^2\text{varA} + c_{11}^2\text{varC} & a_{11}^2\text{varA} + c_{11}^2\text{varC} + e_{11}^2\text{varE} \end{bmatrix}$$

$=$

		Twin1		Twin2	
		P1	P2	P1	P2
Twin1	P1	$a_{11}^2 + c_{11}^2 + e_{11}^2$	$a_{11} \times a_{21} + c_{11} \times c_{21} + e_{11} \times e_{21}$	$1.0/0.5 \times a_{11}^2 + c_{11}^2$	$1.0/0.5 \times a_{11} \times a_{21} + c_{11} \times c_{21}$
	P2	$a_{11} \times a_{21} + c_{11} \times c_{21} + e_{11} \times e_{21}$	$a_{21}^2 + a_{22}^2 + c_{21}^2 + c_{22}^2 + e_{21}^2 + e_{22}^2$	$1.0/0.5 \times a_{11} \times a_{21} + c_{11} \times c_{21}$	$1.0/0.5 \times a_{22}^2 + 1.0/0.5 \times a_{21}^2 + c_{22}^2 + c_{21}^2$
Twin2	P1	$1.0/0.5 \times a_{11}^2 + c_{11}^2$	$1.0/0.5 \times a_{11} \times a_{21} + c_{11} \times c_{21}$	$a_{11}^2 + c_{11}^2 + e_{11}^2$	$a_{11} \times a_{21} + c_{11} \times c_{21} + e_{11} \times e_{21}$
	P2	$1.0/0.5 \times a_{11} \times a_{21} + c_{11} \times c_{21}$	$1.0/0.5 \times a_{22}^2 + 1.0/0.5 \times a_{21}^2 + c_{22}^2 + c_{21}^2$	$a_{11} \times a_{21} + c_{11} \times c_{21} + e_{11} \times e_{21}$	$a_{21}^2 + a_{22}^2 + c_{21}^2 + c_{22}^2 + e_{21}^2 + e_{22}^2$

Figure 1.5: Covariance structure for bivariate quantitative genetic analysis under classical twin design. For brevity the different coefficient values for the additive genetic covariance between MZ (*i.e.*, 1.0) and DZ (*i.e.*, 0.5) co-twins are indicated in the same covariance structure here. P1 and P2, phenotype (trait) 1 and 2, respectively; Twin1 and Twin2, first and second members of twin pairs, respectively. $\Sigma_{MZ/DZ}$, expected covariance structure for MZ or DZ twin pairs. A similar coloring as in Figure 4 of[29] is used here, *i.e.*, cells representing within- and cross-twin (co)variances are colored light and dark grey, respectively.

components are computed by Cholesky 'composition'[b] as follows:

$$varA = a * a' = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} * \begin{bmatrix} a_{11} & a_{21} \\ 0 & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11}^2 + 0 \times 0 & a_{11} \times a_{21} + 0 \times a_{22} \\ a_{21} \times a_{11} + a_{22} \times 0 & a_{21}^2 + a_{22}^2 \end{bmatrix} \quad (1.4)$$

In concordance with the labels for the path coefficients in the path diagram displayed in Figure 1.4, the element labeled "a_{11}" in the matrix "a" in equation 1.4 represents the specific influence of the latent additive genetic factor "A1" on the variance of the first phenotype (labeled "P1" in Figures 1.4 and 1.5). Analogously, the element "a_{22}" in the matrix "a" in equation 1.4 represents the specific influence of the latent additive genetic factor "A2" on the variance of the second phenotype (labeled "P2" in Figures 1.4 and 1.5). The additional element of information provided by this bivariate quantitative genetic analysis with respect to univariate analysis, is provided by the element labeled "a_{21}" (and "c_{21}", "e_{21}" *etc.* for any analogous matrices "c", "e", *etc.*) in the matrix "a" in equation 1.4. This element namely represents the additive genetic component of the covariance between both phenotypes that are included in the analysis. A high estimated genetic correlation (additive genetic covariance component standardized by the pooled additive genetic variance for both traits) suggests the presence of one common genetic factor causing correlation between the values for both phenotypes in different persons. It is important to note that a high genetic correlation among traits does not necessarily mean that genetic factors are important causes of the phenotypic covariation among traits; this is only the case if the traits under consideration are highly heritable.[38]

In this thesis, maximum likelihood-based SEM analysis of cross-sectional data obtained in twin families of the same age is considered. However, SEM can also be used to elucidate the causes of phenotypic change over time, using either longitudinal data obtained in the same individuals or cross-sectional data from individuals of different ages.[27,38,39] The SEM-based quantitative genetic analyses presented in this thesis, conducted on the basis of phenotypic data obtained in a genetically informative sample of individuals, are informative of the relative contributions of genetic and environmental variation to phenotypic differences in a population sample of individuals; the results of these analyses are not informative of the causes of particular values being observed for traits in one individual. Also, it is important to keep in mind that with SEM on the basis of covariance structures, we aim to elucidate the relative contributions of sources of phenotypic *variance*; therefore, the causes for the observation in a population sample of particular *mean* values for traits can not be elucidated with this method,[27] but see reference[40].

[b] It is customary to refer to this type of genetic model as being specified by "Cholesky decomposition";[29,36] however because actually a positive definite covariance component matrix is composed rather than *de*composed, the term "Cholesky composition" might be more appropriate and is therefore used throughout this thesis to denote this type of genetic model.

1.4.2 Hierarchical clustering analysis

Hierarchical clustering analysis is a method to "find groups in data".[41] In other words, it can be used to group (cluster) objects (for example, study participants) or variables (for example, metabolites) on the basis of their relative (dis)similarities, such that objects or variables in the same cluster are more similar to each other than to objects or variables in other clusters.[42,43] Hierarchical clustering analysis is typically applied to multivariate data, *i.e.* when two or more variables have been measured for all objects.

An example of such a multivariate data matrix is depicted schematically in Figure 1.6A. Note that in typical metabolomics data, the number of variables is much larger than in the example presented here. The rows of this data table represent the objects; the columns represent the variables. Thus, in hierarchical clustering analysis the aim may either be to find groupings in the "rows" of the data matrix, or to find groupings in the "columns". Methods for clustering *both* rows and columns also exist (for an overview, see *e.g.*[42]), but these are not considered in this thesis. In the following example hierarchical clustering of objects is considered, but note that the methodology to cluster *variables* on the basis of the data for all *objects* is similar to the methodology to cluster *objects* on the basis of the data for all *variables*.

First, on the basis of the data presented in Figure 1.6A a matrix (Figure 1.6B) of the pairwise distances among the objects in the variable space is computed. In this case, Euclidean distance was chosen as a distance measure but there are other possibilities.[41] For example, the Euclidean distance between the objects labeled "1" and "2" in Figure 1.6A is computed as follows:

$$\delta(\text{“1”}, \text{“2”}) = \sqrt{(6.5 - 2.75)^2 + (2.4 - 6.2)^2} = 5.34 \tag{1.5}$$

The distance matrix is a square symmetric matrix having as many rows and as many columns as there are objects in the original data matrix.

Then, based on the distance matrix, a chosen hierarchical clustering algorithm groups the objects. For this example, the "single linkage" clustering algorithm was chosen, but depending on the problem other algorithms might be more appropriate.[41] The result of this grouping is usually depicted as a so-called dendrogram (right hand side of Figure 1.6 C–E). However, for the purpose of clarity, in the left hand side of Figure 1.6 C–E the grouping of objects in each cluster is depicted in so-called "loop plots", which are basically scatter plots of the data with the clustering indicated.[41] First, the pair of objects that have the smallest Euclidean distance to each other are grouped (Figure 1.6C); in this example these are the objects labeled "2" and "4". Then, the clustering algorithm searches for the 'clusters' (actual clusters or individual objects) that are now the closest to each other (objects "5" and "8", Figure 1.6D). This process is continued until all clusters form one big cluster (Figure 1.6E).

In this thesis, hierarchical clustering is applied for quantitative genetic analysis in two ways. First, in Chapters 2 and 4, it is used to group study participants (MZ and DZ twins) and their nontwin siblings on the basis of their

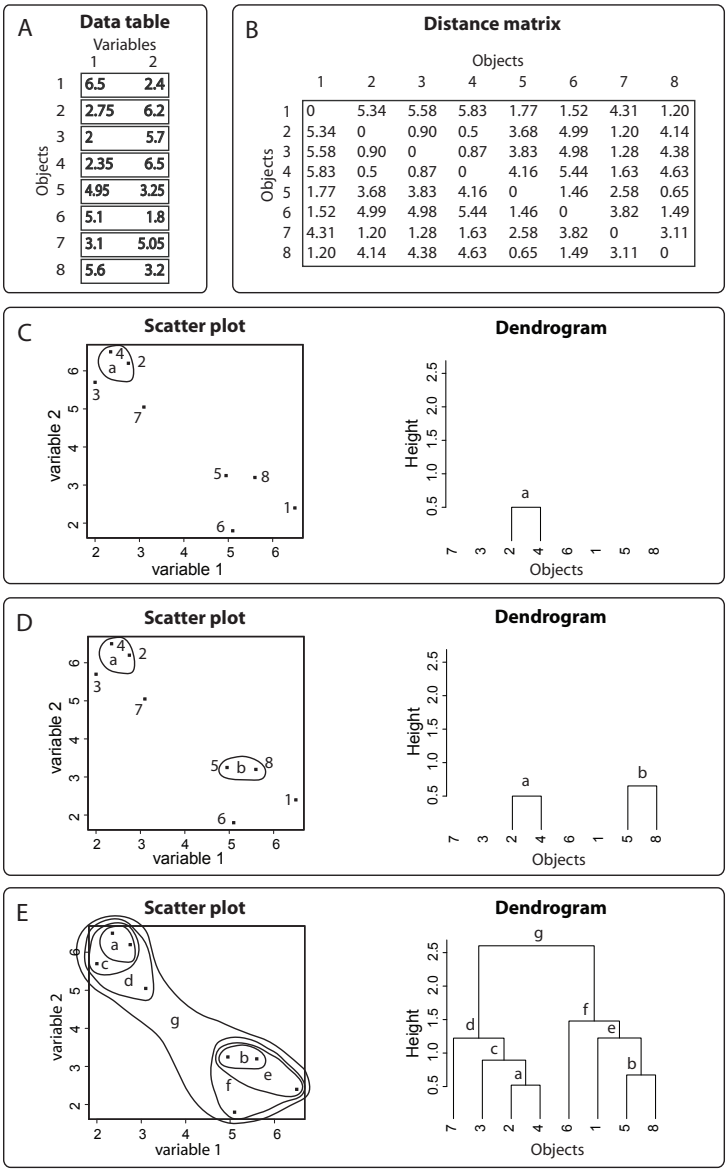

Figure 1.6: Example of hierarchical clustering analysis of eight objects in a bivariate data matrix. The "height" of the branching points in the dendrogram (C–E) equals the distances among clusters as computed by the hierarchical clustering algorithm on the basis of the distance matrix (B). These branching points in the dendrograms have rotational freedom along their vertical axis. Clusters are indicated by alphabetical letters (a–g) both in the scatter or "loop plots" (left hand side in Panels C–E) and in the dendrograms (right hand side of Panels C–E). With kind courtesy of prof.dr.ir. MJT Reinders.

relative similarities in lipidomics profiles. Hence, in this case we are clustering objects (study participants) on the basis of variables (metabolites), analogous to the situation presented as the example in Figure 1.6. Because of the similarities among members of the same family with respect to genetic and environmental effects that might influence metabolite concentrations in body fluids, it is expected that relatives will have relatively similar blood plasma lipidomics profiles and therefore will tend to cluster in hierarchical clustering analysis. On the other hand, it is expected that participants from different families will be put into different clusters by the clustering algorithm because these participants share less genetic and environmental variables than do relatives. Also, it is expected that MZ co-twins will have more similar lipid profiles than DZ co-twins, because of the larger proportion of genetic variance shared by members of MZ twin pairs than by members of DZ twin pairs. In Chapters 2 and 4 of this thesis, it is investigated whether the data provide indications that indeed genetic and environmental similarities among individuals give rise to relatively similar blood plasma lipidomics profiles.

As a second application for quantitative genetic analysis, in Chapter 5 of this thesis hierarchical clustering is used to group variables (metabolites) on the basis of their genetic correlations (see the preceding section of this Chapter for an explanation of the estimation of genetic correlations in multivariate quantitative genetic analysis by SEM). Hence, this type of analysis aims to highlight groups of variables in the data that share genetic causes of their phenotypic (co)variance.

In contrast to in SEM, in hierarchical clustering analysis no explicit model is specified relating predicted variables to predictor variables and tested against measured data; rather, in hierarchical clustering analysis "we just want to see what the data are trying to tell us".[41] For example, when performing cluster analysis of metabolite profiles obtained in a group of individuals (Chapters 2 and 4 of this thesis), we do not specify a model that relates the grouping of participants to the influence of genetic and environmental latent factors. And when performing cluster analysis of dissimilarities computed from "genetic correlations" among different metabolites (Chapter 5 of this thesis), we do not specify a model that relates the grouping of metabolites to the degree to which the concentrations of these metabolites in body fluids are influenced by the same genes.

1.5 Quantitative genetic analysis for systems biology

Systems biology is the study of biology as a holistic system of genetic, genomic, protein, metabolite, cellular, and pathway events that are in flux and interdependent.[44] The interdependence among the elements (such as proteins, metabolites and gene transcripts) of biological systems in quantitative terms can be represented as a correlation network. Such a correlation network is de-

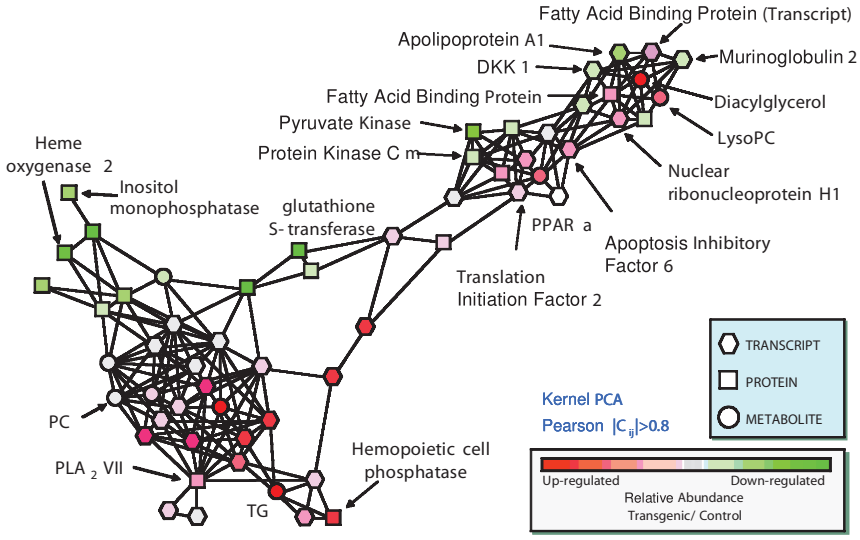

Figure 1.7: Correlation network of select genes, proteins and lipids in the APOE*3 Leiden mouse. The colors of the elements of the system (transcripts, proteins and metabolites) indicate the relative expression in the transgenic animals compared to wild type controls (red=higher level; green=lower level) and a line connecting two elements indicates an absolute phenotypic (Pearson) correlation with a value larger than 0.8. Reprinted with permission from OMICS: A Journal of Integrative Biology Volume 8, Issue 1, 2004, published by Mary Ann Liebert, Inc., New Rochelle, NY.

picted graphically in Figure 1.7, which was based on the results of integrative (systems biology) analysis of the APOE*3 Leiden mouse. [45]

The APOE*3 Leiden mouse is a transgenic animal that differs from wildtype animals because it expresses the human APOE*3 Leiden gene. The importance of this mutation for the relative concentrations of elements in the system in the transgenic mouse with respect to the wildtype mouse is indicated in Figure 1.7 with the color of the elements. That is, elements that are colored red or green in Figure 1.7 are heavily influenced by the genetic difference between wildtype and APOE*3 mice, whereas no quantitative influence was observed of this genetic factor on the colorless elements. In quantitative genetic terminology, one might state that the "heritability" of the elements colored red or green in Figure 1.7 is high (under the given condition of equal environments for wildtype and transgenic animals). Also, in Figure 1.7 elements of the system that have a high phenotypic correlation with each other in both the wildtype and in the transgenic animal are connected by lines.

A phenotypic correlation network such as in Figure 1.7 could be generated for humans as well on the basis of structural equation modeling, if transcriptomics, proteomics and metabolomics data have been obtained in a genetically

informative sample. However, as SEM-based multivariate analyses are also informative of the genetic and environmental structure underlying phenotypic correlations, such analyses would in addition allow the 'decomposition' of the phenotypic correlation network into a 'genetic correlation network' and an 'environmental correlation network'. Also, the elements (*i.e.*, transcripts, proteins and metabolites) of the phenotypic correlation network would look different from those in Figure 1.7. That is, because quantitative genetic analyses are typically based upon a *polygenic*[46,47] rather than a *monogenic* model of genetic variation (as could be conceived for the APOE*3 Leiden mouse), the phenotypic network could indicate to what extent the elements vary quantitatively in a *continuous* manner rather than in a dichotomous manner due to genetic variation. Also, probably not all elements showing the same degree of heritability could be given the same color, as in Figure 1.7, because different genes might influence different elements. Genetic correlations as estimated with multivariate quantitative genetic analyses on the basis of SEM could indicate to what extent the latter is indeed the case.

1.6 The value of our approach in the (post-)GWA study era

The quantitative genetic analyses as described in this thesis use phenotypic data obtained in a genetically informative sample of individuals to partition phenotypic variation into genetic variation and environmental variation. In such analyses on the basis of SEM, the "genetic factors" and "environmental factors" are modeled as latent variables (*e.g.*, "$A1$" and "$C1$", respectively). However, arguably, the genome-wide association (GWA) study is currently the most widely used technique for studying the association between DNA sequence variation (in the form of single-nucleotide polymorphisms, SNPs) across the genome and variation at the phenotypic level (degree of expression of qualitative or quantitative traits).[48] A typical GWA study provides statistical significance values of the associations between variation in each SNP or haplotype, and variation in each phenotype of interest for a particular cohort.[49,50] The results of GWA studies that consider quantitative traits can be used to model the dependency of trait values on the allele copy number of significantly associated SNPs/haplotypes. Thereby, such GWA studies also provide the proportion of phenotypic variance explained by a particular SNP ([47]; for examples, see[12,51]). Therefore, in contrast to the type of quantitative genetic studies described in this thesis, in GWA studies the measurable or 'manifest' genotypic variables (*i.e.*, SNPs indicating quantitative trait loci) are elucidated that influence variation in a trait. Recently, GWA studies have been performed linking genomic variation and variation in metabolomics data.[12,51,52]

It is argued here that the type of quantitative genetic studies as described in this thesis are able to provide valuable information in the context of the recent developments in GWA studies. Specifically, heritability estimates, as provided

for example by twin or family studies on the basis of phenotypic data, can provide a reference point for the proportion of phenotypic variance explained by SNPs that are significantly associated with the phenotype.[53–56] In contrast to most GWA studies, such twin or family studies acknowledge that an in theory infinite number of *polygenes* contributes to phenotypic variation, without making inference on the identity of these genes.[47] If, for example, the total proportion of phenotypic variance explained by all significantly associated SNPs in a GWA study is notably smaller than the heritability as estimated using *e.g.* twin studies, then this is often called "missing heritability".[48,57,58] Indeed, the concept of 'missing heritability' has caused a ripple in the recent literature,[49,58–60] notably because for common diseases (and common traits, like height) GWA studies have not been able to explain much heritability yet[57,58] although novel analysis strategies for GWA bear much promise.[61,62] Common traits/diseases, which are presumably influenced by a large number of polygenes and a large number of environmental factors, are becoming increasingly important as study objects.[1] Oft-mentioned potential causes for missing heritability in GWA studies for common traits are, amongst others, that common genotypes of small effect size or rare variants are important contributors to heritability but are missed by GWA studies.[48,63,64] Recently, within the context of diseases, this view was refuted by Clarke and Cooper,[65] who stated that 'missing heritability', especially for diseases that are relatively severe, might be explained by natural selection on the basis of *de novo* genetic variation, which is not detected by GWA studies. However, for quantitative traits, as well as for common diseases that might well just represent the extremes of the distributions of a number of quantitative traits,[47,66] this latter view might not be applicable as such quantitative traits will not be subject to stringent natural selection.[67]

Rather than to compare the proportions of phenotypic variance attributable to 'genetic variation' and particular SNPs in quantitative genetic studies and GWA studies, respectively, heritability estimates can also be used prior to embarking on GWA studies to estimate the likelihood that genotypes associated with the phenotypes of interest will be detected with statistical significance.[37] Furthermore, the detection of pleiotropy (shared set of genes influencing multiple traits) by multivariate quantitative genetic analyses can support observations in GWA studies of statistically significant associations of different traits with the same SNPs. Because of similar reasoning, the type of quantitative genetic analyses described in this thesis might contribute to the interpretation of the findings from *e.g.* "gene-environment-wide interaction studies" as well.[68]

1.7 Outline of this thesis

In Chapter 2, hierarchical clustering analysis is used to cluster members of (mainly MZ) twin families on the basis of their blood plasma lipidomics profiles. The results suggest an important role for genetic effects and for gender in

determining similarities of lipidomics profiles among individuals. Also, the results suggest that lipid profiling might be useful for monitoring personal health, because dissimilarity of blood plasma lipidomics profiles in a number of cases corresponded both with increased levels of the acute inflammatory marker C-reactive protein (CRP) and with self-reported recent illness.

The *power* of any statistical analysis will be enhanced by increasing the number of observations, and this holds in particular for twin studies.[69] Therefore, in Chapter 3 of this thesis, a data pretreatment method is described to make combinable metabolomics data sets obtained with the same analytical method but on different occasions. The application of this method is demonstrated with data sets obtained by ^{1}H NMR spectroscopic analysis of blood plasma and of urine, and by LC–MS analysis of blood plasma lipids.

In Chapter 4, the method to cluster family members on the basis of their lipidomics profiles as presented in Chapter 2 is applied to the combined LC–MS data sets obtained after application of the method presented in Chapter 3. The combined data set contains data for MZ as well as DZ twin families. Hierarchical clustering analysis of these data supported the finding in Chapter 2 of the relatively large contribution of shared genetic background to similarity of lipidomics profiles among individuals. In addition, the results suggest that shared environmental influences are also important for such similarity. In line with the findings presented in Chapter 2, female gender correlated positively with dissimilarity of lipid profiles in MZ twin pairs.

In Chapter 5, uni- and multivariate quantitative genetic analyses on the basis of SEM are applied to the blood plasma ^{1}H NMR data set and the blood plasma lipid LC–MS data set obtained by using the data set combination method described in Chapter 3. For the multivariate analyses a "multistep multivariate" method was applied that is demonstrated in this Chapter to be useful for the relatively hypothesis-free analysis of "omics" (such as metabolomics) data sets.

Finally, general conclusions are drawn and future perspectives are discussed in Chapter 6.

CHAPTER 2

Similarities and Differences in Lipidomics Profiles among Healthy Monozygotic Twin Pairs

Harmen H.M. Draisma,[1] Theo H. Reijmers,[1] Ivana Bobeldijk-Pastorova,[2] Jacqueline J. Meulman,[3] G. Frederiek Estourgie-Van Burk,[4,5] Meike Bartels,[5] Raymond Ramaker,[2] Jan van der Greef,[2] Dorret I. Boomsma,[5] and Thomas Hankemeier[1]

OMICS A Journal of Integrative Biology 2008:12(1), 17

[1]Leiden University, LACDR, Leiden, The Netherlands.
[2]TNO Quality of Life, Zeist, The Netherlands.
[3]Leiden University, Mathematical Institute, Leiden, The Netherlands.
[4]Department of Paediatric Endocrinology, Institute for Clinical and Experimental Neurosciences, VU University Medical Center, Amsterdam, The Netherlands.
[5]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands.

2.1 Abstract

Differences in genetic background and/or environmental exposure among individuals are expected to give rise to differences in measurable characteristics, or phenotypes. Consequently, genetic resemblance and similarities in environment should manifest as similarities in phenotypes. The metabolome reflects many of the system properties, and is therefore an important part of the phenotype. Nevertheless, it has not yet been examined to what extent individuals sharing part of their genome and/or environment indeed have similar metabolomes. Here we present the results of hierarchical clustering of blood plasma lipid profile data obtained by liquid chromatography-mass spectrometry from 23 healthy, 18-year-old twin pairs, of which 21 pairs were monozygotic, and 8 of their siblings. For 13 monozygotic twin pairs, within-pair similarities in relative concentrations of the detected lipids were indeed larger than the similarities with any other study participant. We demonstrate such high coclustering to be unexpected on basis of chance. The similarities between dizygotic twins and between nontwin siblings, as well as between nonfamilial participants, were less pronounced. In a number of twin pairs, within-pair dissimilarity of lipid profiles positively correlated with increased blood plasma concentrations of C-reactive protein in one twin. In conclusion, this study demonstrates that in healthy individuals, the individual genetic background contributes to the blood plasma lipid profile. Furthermore, lipid profiling may prove useful in monitoring health status, for example, in the context of personalized medicine.

2.2 Introduction

Differences in genetic makeup and in environmental exposure manifest as differences in measurable characteristics in individuals, that is, as differences in phenotypes.[70] Metabolite profiles are regarded as being an important part of the phenotype.[9] It is currently unknown to what extent an individual's metabolite profile is a function of the genotype and of environmental conditions. If genotype is an important determinant of metabolite profiles, it is expected that biological relatives who share genes and possibly also share environments will show similarities in metabolic profiles.

To explore these issues, we carried out a study in healthy, 18-year-old monozygotic (MZ) twins and their biological siblings. The process of obtaining a comprehensive view of the metabolites in an organism has been termed "metabolomics".[9] In humans, metabolomics strategies are often used to find differences in metabolite profiles between groups having different phenotypes, for example, between groups of healthy and diseased individuals.[71] Indeed, with respect to other 'omes such as the genome, the metabolome might be more informative of the physiological state of an organism. For example, in a study where similarities in gene expression profiles of twins discordant for rheumatoid arthritis were compared to similarities in expression between healthy twins, no

difference was found between the healthy twin pairs and the twin pairs where one twin had the disease.[72]

Perhaps the most widely used techniques to measure a wide range of metabolites in biological samples in metabolomics are nuclear magnetic resonance (NMR) and gas or liquid chromatography coupled to mass spectrometry (GC–MS and LC–MS, respectively). NMR aims at obtaining a picture of the complete metabolite profile of a sample, and thus is able to provide a "global" view of the metabolome. Its sensitivity is typically lower than that of MS-based methods such as LC–MS, though. A "targeted" approach, on the other hand, focuses on analysis of particular classes of metabolites, for example amino acids, sterols, or lipids.[73] An LC–MS platform can be used for both global and targeted approaches,[71] but it is impossible to analyze in one run metabolites that have widely differing physicochemical properties such as different polarities and acid dissociation constants. Disadvantages of gas chromatography when applied in metabolomics studies are that often derivatization is necessary,[74] and that even then, only particular classes of metabolites are measurable.

In this study we have applied LC–MS in a targeted manner to obtain lipid profiles in blood plasma samples from healthy MZ and dizygotic (DZ) twin pairs and their siblings. Previous research in our laboratory using the LC–MS method applied in the current study suggested that family members had relatively similar blood plasma lipid profiles, although strong evidence for this was lacking (unpublished results). Furthermore, lipids are especially interesting metabolites because they are involved in a wide range of physiological processes. For example, triglycerides (TGs) serve as an energy source for the body,[75] as a precursor for cell membrane phospholipids,[76] and in the form of body fat they are important for thermal insulation.[77] Among the lipids, TGs are the most important class into which potentially toxic compounds can be incorporated.[78] Another class of lipids with entirely different functions comprises the lysophosphatidylcholines (LPCs). These can be formed from phosphatidylcholines (PCs) present in low-density lipoprotein, for example, by platelet-activating factor acetylhydrolase (lipoprotein-associated phospholipase A2).[79] The activity of this enzyme may be increased upon proinflammatory stimuli;[80] the formed LPCs can act as a chemoattractant for phagocytes.[81] PCs can also act as fatty acid donor for cholesterol esterification by the LCAT enzyme,[82] and may cause platelet aggregation after their oxidation.[83] Bile is partly comprised of PCs.[84] Furthermore, PCs are precursors of sphingomyelins (SPMs), and share some of their functions with them: lipids from both classes are important structural components of cell membranes[85] and of lipoprotein particles.[86] They are also involved in signal transduction,[87] and are constituents of lung surfactant.[88] Whereas the surface of a lipoprotein partly consists of PCs and SPMs, cholesteryl esters (ChEs) are an integral part of its core.[86] The main biological function of ChEs is that they are precursors of steroids.[89]

Twins are particularly informative study populations because the members of pairs share genetic and environmental influences. MZ co-twins share their

complete or nearly complete DNA sequence. Thus, for any heritable trait, they will show phenotypic resemblance. The more heritable a trait, that is, the larger the influence of additive genetic variation on the phenotype, the larger the resemblance in MZ twins. First-degree relatives such as DZ twins and biological siblings share on average 50% of their segregating genes. Therefore, also for these relatives their phenotypic resemblance is expected to be considerably dependent upon the heritability of the traits under consideration. However, resemblance between relatives who are not MZ twins also depends on the genetic architecture of a trait. For example, if non-additive genetic influences such as dominance or epistasis are of importance, phenotypic resemblance in siblings is expected to be relatively low. If, on the other hand, genetic influences are mainly additive, phenotypic resemblance in DZ twins and siblings will be roughly half of the resemblance in MZ twins. If, next to heritability, the shared family environment —in the literature also referred to as the "common environment" or "family environment" [27]— also contributes to phenotypic resemblance of relatives, then first-degree relatives will approach the resemblance of MZ twins more than is expected on basis of genetic segregation.

In classical twin studies, knowledge about genetic and social relationships among co-twins and siblings reared together is used to impose certain structure upon the measurement data.[27] Uni- and multivariate data are often modeled within the context of genetic covariance structure approaches, using estimation techniques based on maximum likelihood. However, such other approaches require that the number of measured variables is not (much) larger than the number of independent clusters (*e.g.*, twin pairs or families) that take part in the study. Therefore, such techniques have rather limited applicability in typical "omics" studies, where the number of measured characteristics is much larger than the number of individual samples. Such a multi- or megavariate approach is the consequence of the idea that when studying biological systems, multiple rather than individual measured variables will reflect underlying, as such unobserved, phenomena. As an alternative, in the current study we have applied an unsupervised approach that is based upon hierarchical clustering of metabolite profiles to identify biologically relevant subgroups of participants (*i.e.*, twin pairs and families) in the data. With this approach, it is possible to get an impression of the within-family variation in metabolite profiles relative to the between-family variation.

We expected to identify clusters of family members in the data, in those cases where family members share relevant genes and/or environment. Coclustering of twins was evaluated using a permutation test. In instances where co-twins did not cluster closely together, we have attempted to provide explanations for this. Our results suggest an important role of genetic background in the generation of interindividual variation in blood plasma lipid profiles. Moreover, several lipids measured in this study may prove to be appropriate for monitoring health status, for example, in the course of personalized treatment.

2.3 Methods

2.3.1 Participants

Participants were recruited from the Netherlands Twin Register at the Vrije Universiteit (VU) in Amsterdam, The Netherlands.[90] The aim was to recruit MZ twin pairs of approximately 18 years old from a cohort participating in a longitudinal investigation into the heritability of mental and physical development in late puberty.[91]

Near the twins' 18^{th} birthday, the twin pairs and their siblings were invited to take part in the project. Ethical approval was given by the Central Committee on Research Involving Human Subjects in The Netherlands. Informed consent and parental consent, if a sibling was under 18, were obtained. Zygosity was determined for all twin pairs by DNA genotyping ($N = 20$ pairs) or using blood group polymorphisms ($N = 1$ pair).

Between November 2004 and September 2005, all participants came to the VU University in Amsterdam for a physical examination in the morning and neurophysiological assessment in the afternoon. Blood was drawn after overnight fasting during the morning session. In addition, subjects completed a series of questionnaires regarding demographics, problem behavior, health, lifestyle, educational attainment, and other traits. For the current study, we used answers to questions regarding current use of any medication, subjective health up to 1 month prior to blood sampling, current and earlier smoking habits, and whether participants currently lived at their parents' home.

2.3.2 Blood sampling

Female participants reported the day of their menstrual cycle at the time of sampling. To prevent clotting, heparin was used as a coagulant and blood collection tubes (BD Vacutainer Systems, Preanalytical Solutions, Belliver Industrial Estate, Plymouth, UK) were inverted gently immediately after collection. About 20 min later, tubes were put on ice. Approximately 2 h following withdrawal, tubes were centrifuged for 20 min at $2{,}100 \times g$ using a Hettich Rotixa 120R centrifuge (Hettich AG, Bäch, Switzerland). Plasma fractions were then transferred to 500 μL cups and stored at -20°C until analysis. For each included family, samples were obtained from every participant from that family at the same day and processed by the same person. The concentration of C-reactive protein (CRP) was assessed in thoroughly thawed frozen heparin samples.

2.3.3 Sample preparation

From each plasma sample, 10 μL aliquots were taken in duplicate. For quality control purposes a pooled sample consisting of equal amounts of plasma from all study participants was prepared and divided into 10 μL aliquots. These

samples (QC samples) were further treated in the same way as the study samples. The samples were divided into two batches, each batch containing one aliquot of each study sample. After separate randomization of each batch QC samples were inserted following each ninth study sample. Samples were deproteinized by adding 300 μL of isopropanol containing the following internal standards: C17:0 LPC 1 μg/mL, C24:0 PC 1 μg/mL, C17:0 ChE 1 μg/mL, and C51:0 TG 1 μg/mL. In this denomination of lipids, the number of carbon atoms as well as the number of double bonds in the fatty acid, separated by a colon (*e.g.*, C17:0) are followed by the class abbreviation (*e.g.*, LPC). After centrifugation, the clear supernatant was collected and the samples were again stored at −20℃ until analysis.

2.3.4 LC–MS lipid profiling

Lipid extract (10 μL) was analyzed using a TSQ Quantum Discovery Triple Quadropule mass spectrometer (ThermoFinnigan, Breda, The Netherlands), equipped with a Surveyor MS HPLC pump and a Surveyor auto injector. The compounds were separated on an Alltech Prosphere C4 300Å HPLC column (150 × 3.2 mm i.d., 5 μm) (Alltech, Lexington, KY) and a Symmetry 300 C4 guard column (10 × 2.1mm i.d., 3.5 μm) (Waters, Milford, MA) using a methanol/water gradient with ammonium acetate and formic acid. After ionization in electrospray (positive mode) the compounds were detected in full scan mode using a scan range of 300–1100 *m/z*.

2.3.5 Data processing/integration

For all detectable lipids a target list was composed based on retention time and *m/z* ratio and the peaks were integrated using LCQuan V2.0 software. The target table comprised lipids belonging to the following classes: LPC, PC, SPM, ChE, and TG. To correct for differences in extract volumes, injection, and changes in signal of the instrument during analysis, all lipid peaks were normalized using the internal standard of that class. The SPMs were normalized using C24:0 PC.

2.3.6 Assessment of the quality of the data

As a measure of the experimental error induced by variation in the sample pretreatment procedure and variation in the measurements over the total duration of the experiment, for each identified lipid compound the standard deviation of its peak areas in the appropriate reconstructed ion chromatograms of the individual QC samples was computed relative to the averaged peak area over all QC samples (relative standard deviation, RSD).

2.3.7 Statistical analysis

Statistical analyses were carried out in the statistical language and environment R (version 2.2.1)[92] and in MATLAB (version R2006b, The Mathworks, Natick, MA). For each sample the replicate measurements were averaged. The resulting data matrix was autoscaled, rendering the mean of the distribution for each lipid compound zero and its variance around this mean one, with the aim to assign all lipid compounds equal weight in the subsequent hierarchical clustering.[93] Then, each row of the data matrix, corresponding to the averaged profile of one study participant and henceforth denoted as "object", was subjected to standard normal variate scaling (SNV)[94,95] to correct for the interindividual differences in the total lipid signal observed by this method. Euclidean distances were computed to measure the dissimilarities among objects. According to the Young-Householder theorem, SNV (applied to the objects) followed by squared Euclidean distance computation is mathematically equivalent to computing $(1-)$ the correlation among unscaled objects.[96]

To assess whether there were differences in median Euclidean distance among (1) MZ co-twins, (2) MZ twins and their same-sex siblings, and (3) same-sex nonfamilial study participants, we performed a multiple comparison procedure using a Tukey's honestly significant difference criterion type of critical value on basis of the result of a nonparametric analysis of the variance within these groups of study participants versus the variance between groups.[97] A multiple comparison procedure is designed to be conservative when testing for significant differences among pairs of groups.[98]

Subsequently, the calculated distances among all objects were subjected to hierarchical clustering analysis. In our choice of the used clustering algorithm we strived for maximum correlation of the distances among clusters as computed by the clustering algorithm (cophenetic distances),[99] with the original Euclidean distances among objects. Of the evaluated clustering algorithms, average linkage gave the highest Pearson correlation (0.71) between the Euclidean distances among objects and the cophenetic distances among clusters, and was therefore considered appropriate. Average linkage minimizes the average of the pairwise distances between objects in different clusters.[100]

To assess the stability of the clustering, we calculated bootstrap probability values (BP values) for each cluster using the R package pvclust[101] and performing 10,000 resamplings of the variables over all objects.

The number of nodes, or branching points, in the resulting dendrogram along the path separating co-twins was then used as a measure of coclustering of co-twins (see Fig. 2.3A for an example). For each number of nodes separating co-twins, we compared the number of observations in the original clustering dendrogram with artificial situations where there is no clustering. In a dendrogram, the "root" of the tree (for example, in Fig. 2.2, the "top" of the dendrogram) is where all clusters ultimately merge, whereas each of the "leaves" at the "bottom" of the dendrogram corresponds to a single object, which is in our study a scaled average lipid profile of one individual. We

created artificial negative control situations by 1000,000-fold Monte Carlo resampling of the object labels over the leaves of the observed clustering tree. For each of the individual permutations, the number of occasions where co-twins were separated by a given number of nodes in the clustering dendrogram was recorded. When the observed number of occasions where co-twins were separated by a given number of nodes in the dendrogram was above the 95% level of the distribution for that number of nodes as resulting from all permutation tests, the observed number of occasions for that number of separating nodes was considered statistically significant.

Based upon this analysis, two subgroups of twins were identified, clustering either closely or not closely with their co-twin. For each case where co-twins did not cluster closely, we evaluated several participant characteristics and environmental factors that could provide an explanation for this.

2.4 Results

2.4.1 Participants

The total study cohort consisted of 54 participants from 23 families (30 males and 24 females), where 24 participants belonged to MZ male twin pairs (MZM) and 18 to MZ female twin pairs (MZF). One male-male and one female-female pair who were found to be DZ after additional genotyping (DZM and DZF; encoded as R_□ and F_○; see the legend to Fig. 2.2 for denotation of individual participants) were also included in the study. From seven families, a twin pair and a sibling of the same sex participated (three MZM, one DZM, two MZF, and one DZF). In one additional MZM family a female sibling (H_●) participated. The average age of the twins was 18.0 years (SD 0.2) and of the siblings 17.4 years (SD 4.3).

According to the interviews, all study participants except H_● lived at home with their parents at the time of the study. Four participants used medication, that is, one twin pair (A_○) used the analgesic/antipyretic Ascal, participant S_○_2 used fluoxetine prescribed for depression, and participant H_● used Marcoumar after a lung embolism. Six participants (F_●_3; I_□_1; I_□_2; M_○_2; T_□_1; and T_□_2) smoked at the time of sampling and two participants (E_○_1 and M_○_1) had smoked in the past. Eight twins (A_○_1; A_○_2; I_□_2; K_□_2; M_○_2; T_□_2; W_□_1; and W_□_2) had had something to eat during the fasting period. In the blood samples of twins A_○_1 and X_-□_2, hemolysis had occurred.

2.4.2 Lipid profiling and data processing

Blood plasma samples were analyzed with LC–MS, yielding profiles of 61 individual lipids per sample, which are listed in Figure 2.3C. The RSDs of the internal standard-corrected responses for the individual lipids in the quality

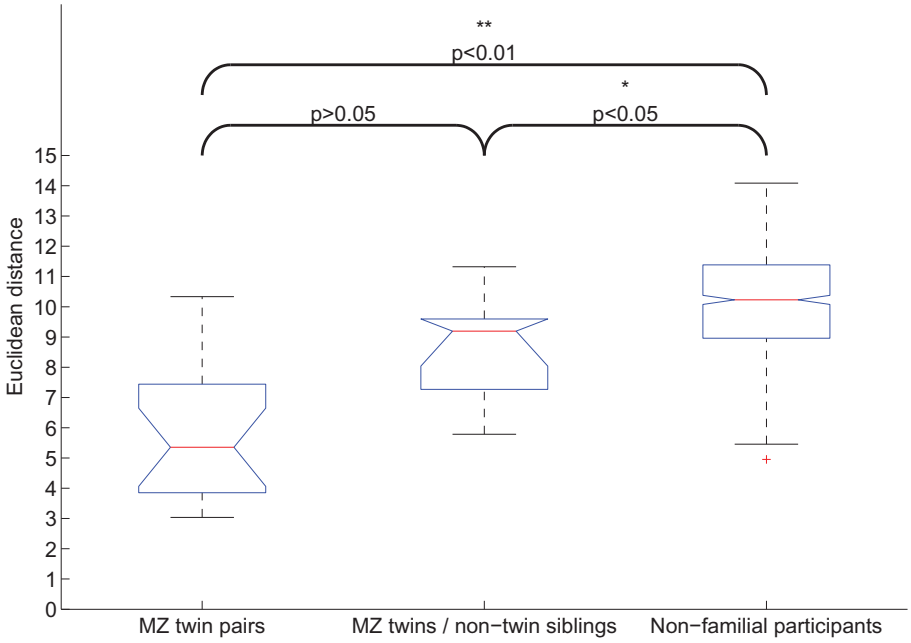

Figure 2.1: Box-whisker plots showing distributions of Euclidean distances between MZ co-twins ($N = 38$, left), between MZ twins and their same-sex nontwin siblings ($N = 10$, middle), and among same-sex nonfamilial participants ($N = 637$, right). Data from co-twins of twins in whose blood plasma samples we had noticed hemolysis (A_○_1 and X_□_2), as well as data from two DZ twin pairs (F_○ and R_□) were included in the computation of distances among nonfamilial subjects only. p-Values as resulting from a multiple comparison test of the group medians are displayed.
$\star p < 0.05; \star\star p < 0.01$.

control samples ranged from 5.2% to 25.5%. Notably, the RSDs of all LPCs, PCs, and SPMs were below 15%.

2.4.3 Statistical analysis

After averaging of the analytical duplicates and scaling of the data table, Euclidean distances between the 54 rows (objects) were computed. The median within-pair distance for MZ twins was significantly smaller than the median distance among nonfamilial participants; the median distance between twins and their same-sex nontwin siblings was also significantly smaller than the median distance among same-sex nonfamilial participants (Fig. 2.1). Similar differences have been observed by Nanki and colleagues for gene expression, which, compared to the metabolome, is of course expected to correlate more strongly with genotype because it is less subject to environmental influences.[72] For the

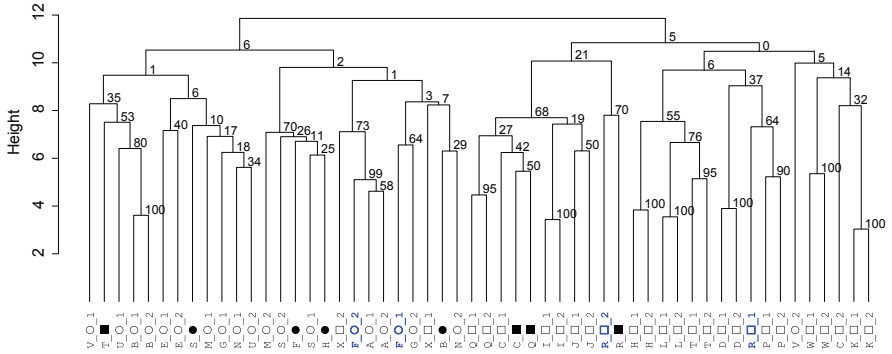

Figure 2.2: Result of nonparametric bootstrap procedure. Numbers near the branching points in the dendrogram indicate BP values on basis of the data resampling method as explained. Denotation of participants: family, alphabetical letter (A-X); sex, squares for males and circles for females (*e.g.*, [□] and [○] for a male and a female twin, respectively); 1,2, randomly allocated to individuals of a twin pair. Labels in bold type indicate DZ twin pairs. Nontwin siblings are indicated by filled squares (■) or filled circles (●) for males and females, respectively.

three investigated gene families, they found that the similarities in expression in peripheral blood lymphocytes were higher between MZ co-twins than among nonfamilial participants.

The result of hierarchical clustering can be displayed as a tree, or dendrogram (Fig. 2.2) that denotes the relationships among clusters in a two-dimensional form. Female and male study participants are almost perfectly separated at the highest level. The dendrogram demonstrates considerable coclustering of MZ twin pairs. However, both DZ twin pairs do not cluster adjacently. Most nontwin siblings do not cluster closely with a sibling who is member of a twin pair. There appear to be rather few clusters that are either extremely tight or extremely loose. Figure 2.3A indicates with a color code Euclidean distances between all pairs of objects. The strong clustering of female (the upper left quadrant in Fig. 2.3A) and of male study participants (the lower right quadrant in Fig. 2.3A) is evident.

In Figure 2.3C, the scaled data is shown for every participant as a separate vertical lane of the heatmap. The order of the objects along the horizontal axis is equal to that in Figure 2.3A. Again, panel C indicates that lipid profiles are different for males and females. The five lipid classes each coincide with a distinct pattern in the heatmap when viewing across all participants from top to bottom. Furthermore, this panel suggests that in general the TGs differentiate less than the other classes between samples from different families. LPCs and SPMs seem to differentiate most. Interestingly, there seem to be differences between families regarding the specific lipid compounds which are most similar among family members.

The stability of the clustering of participants was assessed by a nonparametric bootstrap procedure. For a discussion of the result of this analysis we revert to Figure 2.2. In the context of hierarchical clustering, a bootstrap procedure can be used to investigate to which degree the dendrogram topology changes upon omitting or multiple occurrence of a number of variables for all objects. The stability of the clustering tends to be highest at the lowest level of clustering, that is, where the distance between clusters is relatively small. For the co-twins forming close clusters, BP values were in the range between 40 and 100, and in general clusters containing female co-twins had lower BP values than clusters of male co-twins. Therefore, especially for the female co-twins forming close clusters there may be subsets of variables that are especially important for the clustering. With this in mind, one way to improve the co-clustering of twins may be to use a measure of object similarity, for example, COSA,[43,102] that acknowledges that different subsets of objects may cluster on different subsets of variables.

As a measure of coclustering of co-twins, for each twin pair we counted the number of branching points, or nodes, along the path separating both twins. The colors and heights of the lines that connect twins in Figure 2.3B indicate these numbers. For example, there were three twin pairs where the number of nodes between the co-twins was seven. In the dendrogram of Figure 2.3A an example is drawn of this characterization of the relative similarity of co-twins for a case where the distance between the co-twins is five nodes. For each possible number of nodes separating MZ or DZ twins, the observed frequency is displayed as a black dot in Figure 2.4. Characterizing the clustering of the nontwin siblings with their closest twin brother or sister in a similar way, we found that five of these pairs of family members (*i.e.*, B_●; F_■; H_●; S_●; and T_■ and their closest twin siblings) were separated by more than six nodes, and therefore did not cluster closely. We acknowledge that one difficulty with our approach is that the numbers of branching points along the path separating pairs of objects are not necessarily representative of the absolute magnitude of the dissimilarity between objects, in our case defined by Euclidean distance. For example, co-twins may be dissimilar in terms of Euclidean distance but still be separated by a limited number of nodes. This indicates that although they are dissimilar, they are still more similar to each other than to any other object in their neighborhood within the multidimensional space put up by the lipid profiles of all study participants. Thus characterizing the coclustering of twins in this way gives insight into the similarity of co-twins to each other, relative to the similarity of each individual twin with all other objects in the dataset.

Subsequently, we tested whether coclustering of twins was indeed stronger than what would have been observed by chance, given the observed dendrogram topology. To this end, per possible number of nodes separating twins we created a reference distribution by permutation of the object labels over the leaves of the dendrogram. The significance of the observed numbers of nodes separating twins in the dendrogram was assessed by comparison with

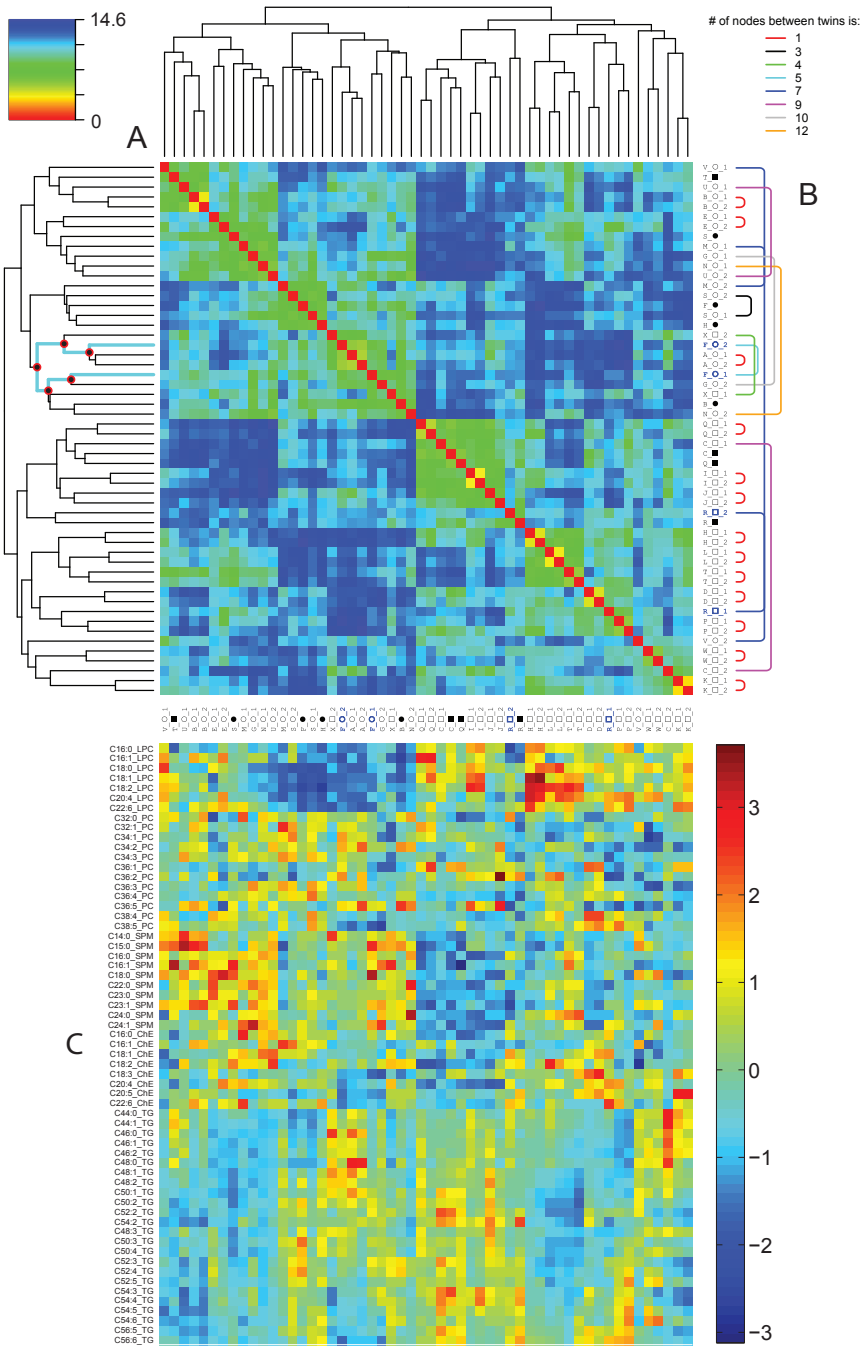
14.6
0
A
# of nodes between twins is:
1
3
4
5
7
9
10
12
B
C
C16:0_LPC
C16:1_LPC
C18:0_LPC
C18:1_LPC
C18:2_LPC
C20:4_LPC
C22:6_LPC
C32:0_PC
C32:1_PC
C34:1_PC
C34:2_PC
C34:3_PC
C36:1_PC
C36:2_PC
C36:3_PC
C36:4_PC
C36:5_PC
C38:4_PC
C38:5_PC
C14:0_SPM
C15:0_SPM
C16:0_SPM
C16:1_SPM
C18:0_SPM
C22:0_SPM
C23:0_SPM
C23:1_SPM
C24:0_SPM
C24:1_SPM
C16:0_ChE
C16:1_ChE
C18:1_ChE
C18:2_ChE
C18:3_ChE
C20:4_ChE
C20:5_ChE
C22:6_ChE
C44:0_TG
C44:1_TG
C46:0_TG
C46:1_TG
C46:2_TG
C48:0_TG
C48:1_TG
C48:2_TG
C50:1_TG
C50:2_TG
C52:2_TG
C54:2_TG
C48:3_TG
C50:3_TG
C50:4_TG
C52:3_TG
C52:4_TG
C52:5_TG
C54:3_TG
C54:4_TG
C54:5_TG
C54:6_TG
C56:5_TG
C56:6_TG
3
2
1
0
−1
−2
−3

Figure 2.3: Euclidean distances among objects and corresponding dendrogram (A); scaled data for each participant (C). In panel B, co-twins are connected by colored lines. In the dendrogram of panel A an example is drawn of our approach to characterize coclustering of twins. The keys to the colors in panels A, B, and C are given in the upper left, upper right, and lower right corners of the figure, respectively. In panel C, lipids are labeled by the number of carbon atoms as well as the number of double bonds (separated by a colon) in the fatty acid, followed by their class abbreviation (LPC, PC, ...).

these reference distributions (Fig. 2.4). In this figure, from left to right, with each separate graph the number of branching points along the path in the dendrogram separating co-twins increases. Due to the given structure of the dendrogram, the maximum possible number of branching points along a path between two leaves was fourteen, and therefore, the number of graphs in Figure 2.4 is also 14. For each number of branching points, from bottom to top the number of twin pairs separated by that particular number of branching points after each permutation is displayed by gray bars. In addition, for each possible number of branching points separating co-twins, the number of twin pairs separated by that number of nodes in the original dendrogram (see Fig. 2.2 and Fig. 2.3A/B) is indicated by black dots. For example, in Figure 2.3A/B, 13 twin pairs can be observed that are separated by only one node. Hence, in Figure 2.4 there is a black dot in the most left graph corresponding with one node separating twins, at the point corresponding with 13 twin pairs. In most permutations, no object labels of co-twins were separated by only one node, and therefore, the horizontal gray bar corresponding with zero observations in the same graph is tallest. As no twin pairs can be observed in Figure 2.3A/B where the number of nodes between twins is two, in the second graph from the left in Figure 2.4 there is a black dot corresponding with zero pairs, and so on.

Using the results of the permutation tests, it was found that the observed number of 13 occasions where co-twins were only separated by one node was significantly different from what would have been observed by chance. For larger numbers of nodes separating co-twins, the observations with the object labels in original order fell within the distributions observed after the permutations. Therefore, we named "close" those co-twins who were separated by one node within the clustering tree and "distant" those co-twins who were separated by more than one node. The notion that there were two subgroups of either close or distant twins in the data, was supported by the observation that the distribution of the within-pair Euclidean distances partly overlapped with the distribution of distances among nonfamilial study participants, as was shown in Figure 2.1.

For each "distant" twin pair, we have attempted to provide an explanation for the observed separation of the co-twins by more than one node in the dendrogram (Table 2.1). These explanations were based upon the available information on participant characteristics and environmental factors. Moreover, in a number of cases dissimilarity of lipid profiles correlated with within-pair differences in the levels of the inflammatory marker CRP (Fig. 2.5). In particular, female sex and recent illness correlated with dissimilarity of lipid profiles between MZ co-twins. In turn, in a number of cases, recent illness as self-reported by the study participants correlated with an increased level of CRP. However, we could not establish the influence of female sex and recent illness independently, because a relatively large number of female study participants had self-reportedly been ill. Moreover, a number of female "distant" twin pairs did not have synchronous menstrual cycles. Dizygosity correlated strongly with dissimilarity of lipid profiles as well, as both DZ twin pairs included in

the study were separated by more than one node in the dendrogram. Moreover, five out of a total of eight nontwin siblings included in the study —of whom all except H_● were of the same sex as their siblings belonging to a pair of twins— did not cluster closely with a sibling belonging to a pair of twins. This observation suggests that the dissimilarity of both DZ twin pairs was caused by differences in genetic background rather than by differences in environmental factors. That is, if shared environment would have been more important for the similarity of lipid profiles, the similarity of nontwin siblings with their twin siblings would have approached the similarity of MZ twin pairs. Although the relative within-pair dissimilarity of lipid profiles correlated with differences in genetic background or environmental exposure, some twin pairs had relatively similar lipid profiles despite the presence of such differences. For example, twins J_□_1 and D_□_1 had had a cold less than 1 week prior to blood sampling whereas their co-twins had not. Still, both discordant pairs were not found to be distant in the clustering. Also, none of the female "close" twin pairs did have completely synchronous menstrual cycles.

2.5 Discussion

In this study we have shown that upon hierarchical clustering of lipid profiles from healthy MZ twins, a significant number of co-twins forms close clusters. This is a strong indication that similarities in genetic background and/or environmental history among individuals indeed manifest as similarities in lipid profiles. Where the genetic resemblance of family members is expected to be lower than in MZ twin pairs on basis of Mendelian inheritance, that is, between DZ twins and between twins and their nontwin siblings, we observed lower similarity of lipid profiles. Moreover, in a number of cases where MZ co-twins did not cluster closely, we have identified recent experiences that might have decreased the within-pair similarity, suggesting an important role of environmental influences in these pairs. Indeed, the similarities among nonfamilial participants, who are expected to share less genetic background and environmental exposure than family members, were low on average.

To our knowledge, this is one of the first reports on unsupervised data analysis of metabolite profiles among healthy twins. Until now, in most publications gene expression data were studied. For example, Tan *et al.*[103] applied their correspondence analysis to project gene expression, measured using whole blood mRNA, separately for each of the 12 elderly female MZ and DZ twins included in their study. They observed that in two MZ and two DZ twin pairs the within-pair correlation of gene expression was higher compared to the correlation between twins from different pairs. Moreover, in these four pairs the within-pair correlation in expression in the MZ pairs was higher than that in the DZ pairs. Omori-Inoue *et al.*[104] performed hierarchical clustering based on correlations between gene expression profiles in umbilical cords from five twin pairs, and found that in four —probably MZ— pairs the co-twins clustered ad-

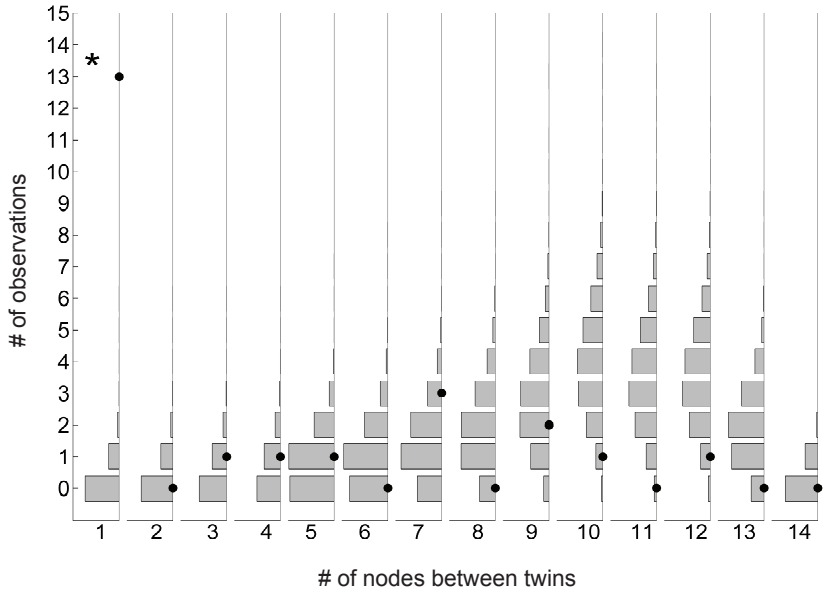

Figure 2.4: Coclustering of twins compared with the results of permutation testing. Numbers of nodes separating co-twins increase from left to right. For each number of branching points, from bottom to top the number of twin pairs separated by that particular number of branching points after each permutation is displayed by gray bars; the number of observations in the original dendrogram (see Fig. 2.2 and Fig. 2.3A/B) is indicated by black dots. The asterisk ($\star$) in the most left graph indicates that the observed number of 13 occasions where co-twins were separated by only one branching point is significantly different from what was observed in the permutation tests.

Table 2.1: Tentative explanations for the separation of co-twins by more than one node in the dendrograms of Figure 2.2 and Figure 2.3A/B

Twin pair	*Explanation*
F_○	(Female) DZ twin pair. F_○_1 had self-reportedly suffered from a cold less than 1 week prior to blood sampling; this correlated with a high blood plasma CRP level in this participant. Both twins used oral contraceptives, but did not have synchronous menstrual cycles.
M_○	M_○_2 had been smoking five cigarettes per day for 6 years and had smoked 2 h before blood sampling; M_○_1 had quit smoking a half year ago after having smoked 10 cigarettes per day for 5 years. Furthermore, M_○_2 had had a half cup of sugared tea for breakfast on the day of blood sampling. Both twins used oral contraceptives, but did not have synchronous menstrual cycles.
R_□	(Male) DZ twin pair. Furthermore R_□_1 had self-reportedly suffered from stomach-ache with cramps less than 1 week before blood sampling.
N_○	N_○_1 had self-reportedly suffered from flu-like symptoms less than 1 week prior to blood sampling; this correlated with an increased blood plasma CRP level in this participant. Both twins used oral contraceptives, but did not have synchronous menstrual cycles.
X_□	X_□_2 had suffered from infectious mononucleosis more than 1 month prior to sampling. Moreover, during sample handling, in the sample of this twin hemolysis had occurred.
G_○	Both twins had self-reportedly suffered from a cold less than 1 week prior to blood sampling. In the blood plasma of G_○_2, a high CRP level was measured.
U_○	Both twins had self-reportedly been ill less than 1 week prior to blood sampling; U_○_1 had suffered from a cold, whereas U_○_2 had had flu-like symptoms accompanied by fever. U_○_2 used oral contraceptives while U_○_1 did not; furthermore, their menstrual cycles were not synchronous.
C_□	C_□_1 had self-reportedly been ill without having a fever less than 1 week prior to blood sampling; this correlated with a high blood plasma CRP level in this participant.
V_○	V_○_1 had reported sickness and headache more than 1 week prior to blood sampling. Both twins used oral contraceptives with synchronous cycles, although V_○_2 appeared to suffer from oligomenorrhea.
S_○	Twin S_○_2 had been using the drug Fluoxetine for depression. Both twins used oral contraceptives, but did not have synchronous menstrual cycles.

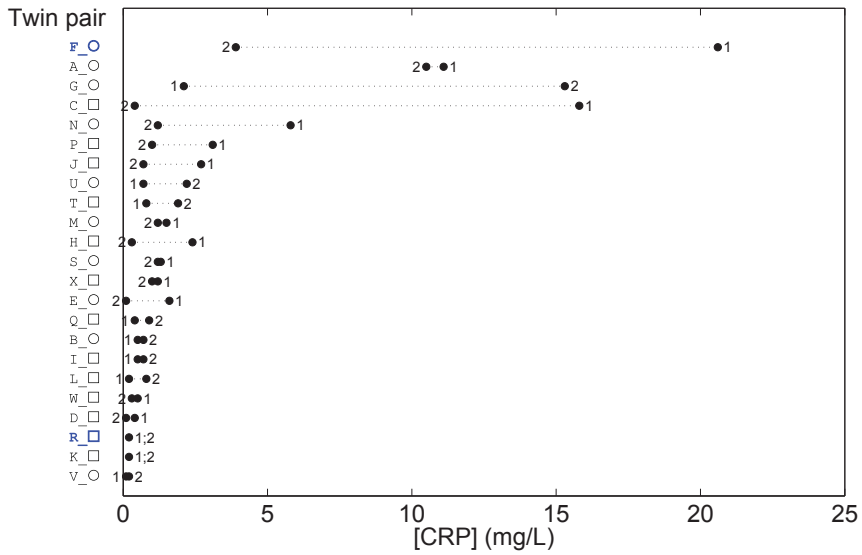

Figure 2.5: CRP levels in blood samples in twins. From bottom to top, the average CRP level of twin pairs increases. The numbers "1" and "2" near the observations denote the "first" and "second" twin of each pair, respectively. For an explanation of this labeling, see the legend of Figure 2.2.

jacently, whereas the non-adjacently clustering co-twins (the fifth pair) might have been DZ. In the result of hierarchical clustering on basis of 102 genes differentially expressed in skin fibroblasts from study participants with systemic sclerosis compared to controls, two MZ twin pairs were observed of which the co-twins —who were discordant for the disease— clustered adjacently.[105] Matigian *et al.*[106] used Pearson correlation as the similarity measure for hierarchical clustering of gene expression profiles in lymphoblastoid cell lines from three MZ twin pairs that were discordant for bipolar disorder, and found that the co-twins of all pairs clustered adjacently. Such high within-pair similarity in MZ pairs was not observed by Teuffel *et al.*,[107] who subjected the bone marrow gene expression profiles of 33 children with acute lymphoblastic leukemia, including one pair of MZ twins concordant for the disease, to hierarchical clustering using Euclidean distances and applying the average linkage clustering algorithm. The authors noticed that the co-twins did not cluster adjacently and ascribed this effect to disease-related changes in gene expression.

Two recent articles employ supervised methods to analyze metabolomics data from twins. In one publication a link was established between schizophrenia and alterations in blood plasma lipid levels as assessed by ^{1}H NMR spectroscopy.[108] Such changes were observed in both male and female affected twins of pairs discordant for schizophrenia when compared to age-matched control twin pairs. However, in females, the differences between the affected twins and their control twins were more pronounced than in males; as opposed to in males, in females the authors also observed a significant difference between the unaffected twins of discordant pairs and control twins. The larger effects in females were attributed to greater genetic predisposition to the disease-related changes in discordant female pairs than in male pairs. A recent study by Pietiläinen and colleagues[109] found within-pair differences in lipid profiles, as assessed in blood serum using LC–MS, in MZ pairs discordant for obesity. Interestingly, the authors report that compared with five normal-weight concordant pairs as well as with five pairs concordant for overweight, the discordant pairs did not have larger intrapair differences in total cholesterol, high-density lipoprotein, low-density lipoprotein or TGs.

Our results suggest that an unsupervised data analysis approach[110] can yield information that can not be derived from other, pseudosupervised analyses. In explorative studies like this one, any constraints in supervised analysis may preclude novel findings. Using hierarchical cluster analysis, we were able to link within-pair dissimilarity of lipid profiles to within twin pair-specific factors.

Studies estimating the relative influence of genetic variation on the within-twin pair variation in lipid levels, have been reviewed by Iselius,[111] and by Snieder *et al.*[112] To our knowledge, with respect to the lipid classes evaluated in this study, only heritability estimates for the TGs have been described previously. A study based upon a population sample having a mean age of 16.7 years, which is close to the mean age of our study cohort, found that genetic factors accounted for 60% of the variation in total TG levels among individuals.[113] In

general, the relative influence of genetic variation on the phenotypic variation in lipid and apolipoprotein levels has been found to be high. Such high estimates are consistent with our findings, of successful clustering of individual twin pairs based upon unsupervised analysis of phenotypic data.

In addition to genetic background and environmental factors, experimental factors such as sample handling and storage, as well as the used analytical methods may introduce further similarities and differences between lipid profiles from different individuals. In our study, lipid profiles were assessed in blood plasma samples from fasting individuals, because during fasting lipid profiles are thought to be relatively stable,[114] and therefore expected to be more similar among individuals sharing genetic makeup and/or environmental exposure. Although we can not completely exclude the possibility that the similarities among samples from MZ twin pairs are partly due to other shared factors induced by the study setup, the larger dissimilarities between twins and their nontwin siblings argue against a strong influence of such factors. Samples from members of the same family (*i.e.*, twins and additional siblings) were collected on the same day. If the workup of samples from a given family would have introduced similarities among the samples from that family relative to samples from other families, a larger resemblance of twin-sibling pairs would have been observed. With respect to sample handling and storage, we suspect that hemolysis of blood samples may augment differences in lipid profiles. In one out of two cases where noticeable hemolysis of the blood sample had occurred, the corresponding twin pair was found to be separated by more than one node in the dendrogram.

We found that healthy individuals who share genetic background and/or environmental exposure, have blood plasma lipid profiles that are more similar than profiles of persons who do not share these influences. When extending this observation in twins to a general healthy population, this probably implies that the lipid profile corresponding with a healthy state is characteristic for each individual due to the individual-specific genetic background and environmental exposure.[73] We therefore suspect that changes in the lipid profile might denote deviations from the healthy phenotype, and therefore could be used, for example, to diagnose the onset of disease. The correlation of an increased blood concentration of the inflammatory marker CRP with dissimilarity of lipid profiles in a number of MZ twins in the current study supports this hypothesis. Actually, it can be assumed that for each individual there is a lipid profile describing the "healthy phenotype", and in the context of personalized medicine the aim could be to maintain this, or to take measures to restore it.

In conclusion, in our study, healthy MZ twins have relatively similar blood plasma lipid profiles. Between individuals with less shared genetic backgrounds and environmental exposure, we indeed observed smaller similarity. Discordance of MZ twins for recent disease, that can be regarded a particularly relevant difference in environmental exposure, correlated well with within-pair dissimilarity of lipid profiles. Therefore, lipid profiling might prove useful in monitoring personal health.

2.6 Acknowledgments

We thank all the participants in this study. We would like to acknowledge support from the Netherlands Bioinformatics Centre (NBIC) through its research program BioRange (project number: SP 3.3.1); Spinoza-premie NWO/SPI 56-464-14192; the Center for Medical Systems Biology (CMSB); Twin-family database for behavior genetics and genomics studies (NWO-MaGW 480-04-004) and NWO-MaGW Vervangingsstudie (NWO number: 400-05-717).

CHAPTER 3

Equating, or Correction for Between-Block Effects with Application to Body Fluid LC–MS and NMR Metabolomics Data Sets

Harmen H.M. Draisma,[1] Theo H. Reijmers,[1] Frans van der Kloet,[1] Ivana Bobeldijk-Pastorova,[2] Elly Spies-Faber,[2] Jack T.W.E. Vogels,[2] Jacqueline J. Meulman,[3] Dorret I. Boomsma,[4] Jan van der Greef,[1] and Thomas Hankemeier[1]

Reproduced with permission from: Draisma, HHM, Reijmers, TH, van der Kloet, F, Bobeldijk-Pastorova, I, Spies-Faber, E, Vogels, JTWE, Meulman, JJ, Boomsma, DI, Van der Greef, J, and Hankemeier, T. Equating, or correction for between-block effects with application to body fluid LC–MS and NMR metabolomics data sets. *Anal.Chem.* 2010:82(3), 1039–1046.

[1]Leiden University, LACDR, Leiden, The Netherlands.
[2]TNO Quality of Life, Zeist, The Netherlands.
[3]Leiden University, Mathematical Institute, Leiden, The Netherlands.
[4]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands.

3.1 Abstract

Combination of data sets from different objects (for example, from two groups of healthy volunteers from the same population) that were measured on a common set of variables (for example, metabolites or peptides) is desirable for statistical analysis in "omics" studies because it increases power. However, this type of combination is not directly possible if nonbiological systematic differences exist among the individual data sets, or "blocks". Such differences can, for example, be due to small analytical changes that are likely to accumulate over large time intervals between blocks of measurements. In this article we present a data transformation method, that we will refer to as "quantile equating", which per variable corrects for linear and nonlinear differences in distribution among blocks of semiquantitative data obtained with the same analytical method. We demonstrate the successful application of the quantile equating method to data obtained on two typical metabolomics platforms, *i.e.*, liquid chromatography–mass spectrometry and nuclear magnetic resonance spectroscopy. We suggest uni- and multivariate methods to evaluate similarities and differences among data blocks before and after quantile equating. In conclusion, we have developed a method to correct for nonbiological systematic differences among semiquantitative data blocks and have demonstrated its successful application to metabolomics data sets.

3.2 Introduction

Combining data from different sources is an important topic in systems biology. At least two types of data combination can be envisaged. The first type of combination is often referred to as data integration or data fusion, and here combination is considered of data sets all representing the same set of objects (for example, a group of healthy volunteers) but different sets of measured variables (for example, metabolites, peptides, *etc.*).[115,116] Data fusion combines the strengths of different analytical techniques to enhance the biological interpretation of the variability present in the study population. In the second type of combination, which is the scope of this article, data sets are combined representing different groups of objects (for example, two groups of healthy volunteers) that were measured on a common set of attributes (for example, the same set of metabolites). Combination of data sets in such a way is desired because it increases the power of statistical analyses. In other words, one may want to combine different data "blocks".

In this article, we use the term "blocks" to refer to measurements obtained on the same analytical method but on different sets of objects and in particular with a considerable time span in between these sets of measurements. A block can consist of data from one or more measurement batches. A similar definition of blocks is given by Zelena *et al.*[117] Different measurement blocks can arise within a study, for example, because (1) the number of study samples is too

large to measure all samples in one measurement block or in one laboratory, (2) additional samples become available in the course of the study while previously collected samples have already been measured, or (3) following a successful pilot experiment, additional samples are measured for validation. It is also conceivable that it is desired to combine data blocks from different studies.

Nonbiological differences between the data from different measurement blocks can exist due to small analytical differences that are often unavoidable and that are typically not addressed during method robustness tests. Such analytical differences are, for example, likely to accumulate over large time spans between blocks of measurements.[117–119]

In data fusion, often three types of combination of data from a common set of objects are considered: high-level fusion, which is the combination of results of data analyses obtained on sets of different variables, low-level fusion, or the concatenation and possibly subsequent weighting of data matrices in such a way that the objects are the shared mode, and mid-level fusion, a term used to describe the combination of variables selected from different data sets.[115,116] A similar classification can be envisioned when considering combination of data on sets of different objects where the attributes are identical. In this article, we present a method that enables such combination of data blocks at a "low level" and illustrate its use with metabolomics data sets. Combination at low level allows maximal flexibility in the choice of subsequently applied (multivariate) data analysis methods yielding results for the combined data sets and therefore is particularly suited to increase the power of such subsequent data analyses. Moreover, combination of data at a low level allows to account for differences in distribution shapes of the same variable(s) among the data sets to be combined, if it is known that such differences have a nonbiological cause. The necessity and possibility of applying data correction methods in order to obtain combinable "omics" data blocks will vary from situation to situation.

In the discussion below, we have intended to provide a guideline where we start with a description of situations where combination should be possible without additional data correction and end with a description of situations where the data transformation method we propose in this article could be useful.

1. If the between-block reproducibility of the used analytical method is good (*e.g.*, semiquantitative nuclear magnetic resonance (NMR) spectroscopy under similar conditions for all measurement blocks of which data sets are to be combined),[120,121] or the data sets to be combined all contain quantitative data (either through separate calibration per measurement block or through transfer of calibration models),[118,119] then the combination of data sets from different measurement blocks should be possible without additional correction. However, currently obtaining quantitative data from metabolomics experiments is still rather difficult, because often due to the absence of reference standards for all detected compounds it is impossible to create a complete calibration model per variable.[122] Both

techniques that are the most frequently used in metabolomics, *i.e.*, liquid chromatography–mass spectrometry (LC–MS) and NMR, suffer from this problem.

2. If the measurements performed within particular blocks are not reliable, then the data from these measurements should be discarded. The reliability of measurements can be monitored using, for example, a quality control (QC) sample consisting of pooled individual study samples, of which aliquots are measured during all analytical measurement blocks.[122–127]

3. Recently, a method has been presented to correct for between-batch effects using these repeated measurements of QC samples as well.[128] Like the other methods to be discussed below, it can be used for the correction of semiquantitative data, *i.e.*, in cases where no full calibration models can be made. We will refer to techniques that make combinable sets of semiquantitative data as "equating" methods, because the term "equating" is used in psychometrics to denote techniques that solve similar problems.[129,130] In the method of van der Kloet *et al.*, the data are corrected for within-batch and between-batch effects per metabolite using the responses of pooled QC samples (for that metabolite).[128] This method can be of use if a single-point calibration is appropriate for correcting differences in data distributions among measurement batches or even among measurement blocks. Of course, it can be used only if the same QC samples are measured in all batches or blocks of which data need to be combined.

4. There are situations where repeated QC sample measurements cannot be used for between-batch effect correction or for between-block effect correction. An obvious example is if such measurements have not been done during all measurement batches or blocks of which data sets need to be combined. Another example is when the QC samples are not representative for the measurements in all data sets to be combined. This can happen for instance if there is differential degradation in the QC samples with respect to the individual study samples. Such situations are analogous to the situations where in the context of multivariate calibration transfer one would typically use "nonstandardization methods", *i.e.*, data preprocessing methods that are independent of transfer standards.[118] An example of an equating method that is independent of repeated QC sample measurements is local autoscaling: autoscaling per data set separately.[131] Like the method described in ref[128], this local autoscaling method could be regarded as a linear equating method.

5. Finally, the data distribution shapes of the same variable in all data sets to be combined can be different mainly due to nonbiological differences among the blocks. Such nonlinear differences among the data distribution shapes in different blocks can arise even if within each block the

measurements for each variable are within the dynamic range of the detector. For example, in case of LC–MS, in a typical metabolomics study, measurement values can be outside the linear range for various reasons: saturation of the detector, peak integration effects (*e.g.*, caused by peak tailing, depending on the concentrations of a particular compound in the samples measured in a particular block), or nonlinear losses during sample preparation. These effects can be different for different measurement blocks. In this article, we propose an equating method that corrects for nonlinear differences between distributions under the assumption that there is an underlying common distribution. Therefore, the beneficial effects of our method will be largest when the compositions of the object groups are balanced among the measurement blocks of which data are to be combined. Our method is independent of repeatedly measured QC samples as well.

In case it has been decided that equating methods need to be considered to correct the data for between-block effects, the choice of a particular equating method might not be trivial. It can be generally stated that the equating method should be used that removes most analytical between-block variation with respect to the biological variation present in all blocks. In practice, however, it is not always possible to determine exactly which part of the total between-block variation is attributable to biological variation and which part is attributable to analytical variation, because the objects measured in different blocks are different. In this respect, an objective evaluation of the results of equating is necessary, because the best equating method in a given situation is not necessarily the one that gives the most desirable results in view of the biological question. Therefore, as with any data preprocessing, using the results of subsequent data analyses alone as a reference to "optimize" the choice for a particular method could lead to bias.

The structure of the remainder of this article is as follows. In the Materials and Methods section, we first introduce the metabolomics data that we will use to illustrate the use of our equating method. Then, we describe our equating method. Univariate as well as multivariate parameters are described that can be used to evaluate the comparability of data sets before and after equating. The Results and Discussion section describes the results of application of our equating method to the data sets originating from the different measurement blocks. Several possible sources of nonbiological systematic variation between data obtained in the different blocks are pointed out. The results of application of our equating procedure to metabolomics data sets, as described in this article, will be used to reproduce and extend our observations that were done in a cohort of twins (see Chapter 2). The results of these subsequent analyses on the combined equated data sets described in the current article will be presented in a separate paper, because the biological interpretation of the results is out of the scope of this paper.

3.3 Materials and methods

Participant recruitment and characterization, blood sampling, and blood plasma sample preparation were performed as described in Chapter 2. In brief, blood was drawn and urine collected from all participants (twins and biological non-twin siblings) after overnight fasting. Plasma samples were stored at −80°C until analysis.

The LC–MS and ^{1}H NMR measurements were performed in two blocks; the measurements of "block 2" (B2) were performed almost 1 year (48 weeks) after those of "block 1" (B1). In B2, for the purpose of QC of the LC–MS and NMR analyses, QC samples were prepared prior to sample preparation by pooling equal amounts of plasma sample from all participants who were measured in that block. In B1, such QC samples were prepared for the LC–MS analyses only. For both LC–MS and NMR analyses, these QC samples were inserted uniformly distributed after separate randomization of the measurement order of the individual study samples in each batch.

3.3.1 LC–MS plasma lipid profiling

Plasma lipid extraction and profiling by LC–MS were performed as described in Chapter 2. After lipid extraction, all extracts were stored at −20°C and measured within 2 weeks. Each peak area obtained for a lipid was corrected using an appropriate internal standard (IS), which had been added prior to sample preparation; no further normalization of the data was applied.

3.3.2 ^{1}H NMR analysis of plasma

Prior to ^{1}H NMR spectroscopic analysis, 300 μL of each plasma sample was centrifuged to remove proteins that had come out of the solution after freezing and transferred to a 5 mm o.d. NMR tube. To each sample 300 μL of deuterated sodium phosphate buffer (0.1 mmol/L, pH 7.4, made up with D_2O) was added.

^{1}H NMR spectra were acquired in triplicate on a fully automated Bruker Avance 600 MHz spectrometer (Bruker Analytik GmbH, Karlsruhe, Germany) using a "Carr-Purcell-Meiboom-Gill" (CPMG) spin-echo pulse sequence and operating at an internal probe temperature of 300 K. The water signal was removed by a presaturation technique in which the water peak was irradiated with a constant frequency during the relaxation delay. A total of 128 transients were acquired into 32×10^3 data points for B1 and 64×10^3 data points for B2. A spectral width of 6 kHz for B1 and 12 kHz for B2 was used with a spin relaxation delay of 88 ms and $\tau 3.4 \times 10^{-4}$ s for both blocks.

The spectra were processed using XWIN-NMR software (v.3.1, Bruker Analytik GmbH). An exponential linebroadening function of 0.5 Hz was applied to the free induction decays (FIDs) prior to Fourier transformation. All spectra were manually phased, baseline-corrected, and referenced to the lactate signal

(CH3 δ 1.33).

After peak picking of the NMR data using the XWIN-NMR software, peak lists were imported into Winlin (V1.10, TNO, The Netherlands). Small variations in chemical shifts in the NMR spectra were adjusted manually based on the partial linear fit algorithm.[132] The peak-picked data from B1 and B2 were aligned together, with the aim to make the alignment for data from both blocks as comparable as possible.

Peaks detected in at least 80% of the spectra recorded in each block were kept for further analysis.[116,127] Then, the data were median-normalized.[133]

3.3.3 Differences between B1 and B2

The 54 healthy participants (30 males and 24 females) who contributed the samples measured in B1 have already been described in Chapter 2. In B2, plasma samples from 128 additional healthy participants (49 males and 79 females) from 42 families were measured. In this cohort, there were 16 monozygotic twin pairs, 26 dizygotic twin pairs, and 44 nontwin siblings. The average age of the twins in the cohort of whom samples were measured in B2 was 18.2 years (standard deviation (SD), 0.2); the average age of the siblings was 19.5 years (SD, 4.8).

In B1, for LC–MS analysis two aliquots were taken of the plasma sample from each individual participant, which were then divided into two measurement batches where each batch contained one aliquot of each study sample. In B2, on the other hand, only one aliquot of each study sample was processed and analyzed in one measurement batch.

Furthermore, following every other of the QC sample aliquots consisting of B2 study samples, aliquots were inserted of the QC sample that had been measured in B1 as well and that thus consisted of B1 individual study sample aliquots (sample pretreatment was performed for this B1 QC sample in B1 and in B2 separately). This B1 QC sample thus underwent an additional freeze-thaw cycle between B1 and B2.

As a measure of experimental error, for each detected lipid compound relative standard deviations (RSDs) were computed for B1 of the IS-corrected measurements in B1 of the pooled QC sample prepared from individual study samples measured in B1, and for B2 of the IS-corrected measurements of the pooled QC sample prepared from samples measured in B2.

In B2, for NMR analysis following each of the QC sample aliquots consisting of B2 study samples, samples were inserted of in total 12 participants that had already been analyzed in B1. These samples thus underwent an additional freeze-thaw cycle between B1 and B2.

3.3.4 Equating data from B1 and B2

Our equating method lets the data for each variable assume the same distribution in all blocks, by averaging the distributions for that variable in all blocks.

An algorithm to achieve this has been presented by Bolstad *et al.*[134,135] This algorithm was based on the principle of the quantile–quantile plot (Q–Q plot). Generally stated, quantiles are the values marking the boundaries between regular intervals of the cumulative distribution of a data sample. That is, when dividing ranked data into a number of subsets, then the quantiles are the values at the boundaries between consecutive subsets. In a Q–Q plot, the quantile values of two distributions are plotted against each other; the number of quantiles plotted equals the number of data points in the smaller data sample (the quantile values in the larger data sample are found by linear interpolation).[136,137] If in the Q–Q plot the points defined by the values of corresponding quantiles in both data samples all lie on a straight diagonal line, then the distributions of both samples are highly similar; if they do not, then the distributions are dissimilar.

In the algorithm as presented by Bolstad *et al.*, the averaging of data distributions is achieved by projecting the corresponding quantile values of all distributions onto a scalar multiple of the unit vector (a, possibly multidimensional, analogue of the diagonal in the Q–Q plot) (Figure 3.1).[134,135] Then, the averaged quantile values are substituted for the original values that are in the subsets belonging to the corresponding quantiles in the data samples under consideration. Thus, the original ranking of the data points in the data samples to be combined is retained. The result is that the distributions of all data samples become equal, or —in the case of different numbers of observations per data sample— almost equal.

This algorithm is usually applied in an "omics" context to make the distributions of different objects equal over all measured variables, that is, for "normalization". Examples of this application are found, *e.g.*, in the fields of genomics (normalization of gene probe intensity distributions between oligomicroarrays, over all gene probes)[135,138–140] and of peptidomics (normalization of peptide intensity distributions between analytical samples, over all detected peptides).[141] However, we introduce the use of this algorithm for equating, that is, for making the distributions of the same variable (NMR feature or lipid) equal over all sets of objects (sets of study samples in all blocks). Because our method is conceptually akin to what is known in psychometrics as "quantile equating" or "equipercentile equating",[130,142] we will refer to it as "quantile equating" as well. Of note, in quantile equating in a psychometrical context the aim is not to make the distributions of the same variable equal for all sets of objects but to provide transformations by which equivalent scores can be found on different versions of the same test.

We used the "normalize.quantiles" function, which was written by the first author of the original publications,[134,135] to perform quantile equating. This function is part of the "preprocessCore" package, which is a component of the Bioconductor software suite (version 2.1)[143] running in the statistical environment `R` (version 2.6.2).[144] For its originally intended purpose, *i.e.*, for normalization, the "normalize.quantiles" function is applied simultaneously to all *objects* (study samples). To perform *equating*, however, we applied this

function to the *variables*. Moreover, we applied the function to the B1 and B2 data for each variable separately.

In case of the LC–MS data, replicate measurements of the individual study samples in B1 were first averaged before equating, whereas in case of the NMR data unaveraged replicates were equated.

Data for samples measured in B1 as well as in B2 (for example, QC samples prepared on basis of pooled aliquots of B1 individual study samples) were omitted from all B2 data sets before equating for the following reason. If the composition of QC samples changes differently between measurement blocks with respect to the composition of individual study samples, then QC samples are not representative for the samples measured in all blocks. In this paper, we show an example of this in case of plasma NMR spectroscopy, where repeatedly measured samples underwent an additional freeze-thaw cycle between B1 and B2 with respect to the individual samples measured in B2. If we would have left the data for these repeatedly measured samples in the B2 block, these data would have influenced the B2 data distributions and thereby would have distorted the result of quantile equating. We did not remove the B1 and B2 measurement data for the QC samples prepared on basis of samples measured in each block, because these helped to visualize the beneficial effects of quantile equating in making combinable B1 and B2 data sets.

3.3.5 Evaluation of comparability of data sets

The comparability of data sets obtained with the same analytical method but in different measurement blocks was evaluated using various methods. At the univariate level, before quantile equating we assessed to which extent the relationship between data distributions of both measurement blocks was nonlinear using the Pearson correlations between the ranked quantile values of both measurement blocks. Due to the nature of quantile equating, after equating the correlations between the B1 and B2 quantile values are always equal to 1.

We characterized the extent to which nonlinear relationships between the distributions as well as other differences between the data from both measurement blocks before equating gave rise to differences at the multivariate level, using a strategy proposed by Jouan-Rimbaud *et al.*[145] In this strategy, data sets are compared in the principal component (PC) space using three continuous parameters that each can take a value between 0 and 1, where a zero value indicates low similarity of the evaluated data sets and a value of 1 suggests perfect similarity. The first parameter ("P") is based upon the comparison of principal components analysis (PCA) loadings patterns, the second parameter ("C") is based upon the comparison of variance-covariance matrices, and the third parameter ("R") characterizes the similarity in location of the centroids of the data sets. The degree of success of quantile equating in making data from both measurement blocks comparable, was characterized using these multivariate parameters as well. We used a 2% increase in total variance explained by the model as a criterion to estimate the number of PCs for which

these parameters were to be computed (PLS_Toolbox version 3.5, Eigenvector Research, Wenatchee, WA).

Furthermore, the success of the equating procedure was visualized by the results of PCA on the combined (concatenated with the variables as the shared mode) data sets originating from different measurement blocks. For this PCA, replicate measurements were averaged. LC–MS data were then mean-centered, whereas NMR data were autoscaled. These different types of scaling were applied to the respective types of data because this enhanced the visibility of the between-block effects prior to equating. All PCA were carried out using the PLS_Toolbox for MATLAB (version R2006b, The Mathworks, Natick, MA).

3.4 Results and discussion

3.4.1 Analytical data

In Chapter 2, the data denoted in the current paper as the B1 LC–MS data have already been presented. The 61 different lipids that were detected in the chromatograms in B1 (see Chapter 2) were detected in B2 as well. Lipids from the following classes were detected: lysophosphatidylcholines (LPC), phosphatidylcholines (PhC), sphingomyelins (SPM), cholesterol esters, and triglycerides (TG). Throughout the manuscript, lipids are denoted as follows: the number of carbon atoms as well as the number of double bonds in the fatty acid, separated by a colon (*e.g.*, C36:5) is followed by the class abbreviation (*e.g.*, PhC).[127] The data for C16:0_LPC and C52:2_TG were excluded from further analysis because their responses displayed a systematic trend in the QC sample measurements in B2, resulting in high RSDs. In B1, the mean RSDs for the remaining 59 lipids as computed on basis of the measurements of the QC sample prepared in B1 were 13.3% (SD, 5.6; range, 5.2–25.5%). Notably, the RSDs of all LPCs, PhCs, and SPMs were below 15%. In B2, the mean RSDs of these same 59 lipids, computed on basis of the measurements of the QC sample prepared in B2, were 7.5% (SD, 1.4; range, 4.9–10.9%). In the plasma NMR data, after application of the "80% rule", 75 features (variables) were kept for analysis.

3.4.2 B1–B2 comparison before equating

PCA scores plots

Panels A and C of Figure 3.2 display the PCA scores plots for the LC–MS and the NMR plasma data, respectively, before equating. As expected, the scores of almost all pooled B1 and B2 QC sample aliquots are in the centers of the clusters corresponding to B1 and B2, respectively. However, in particular in case of the LC–MS data, the scores of the measurements from both blocks display notable separation along the PC1 axis (Figure 3.2A). This phenomenon might have been caused, for example, by slightly different IS concentrations. Another

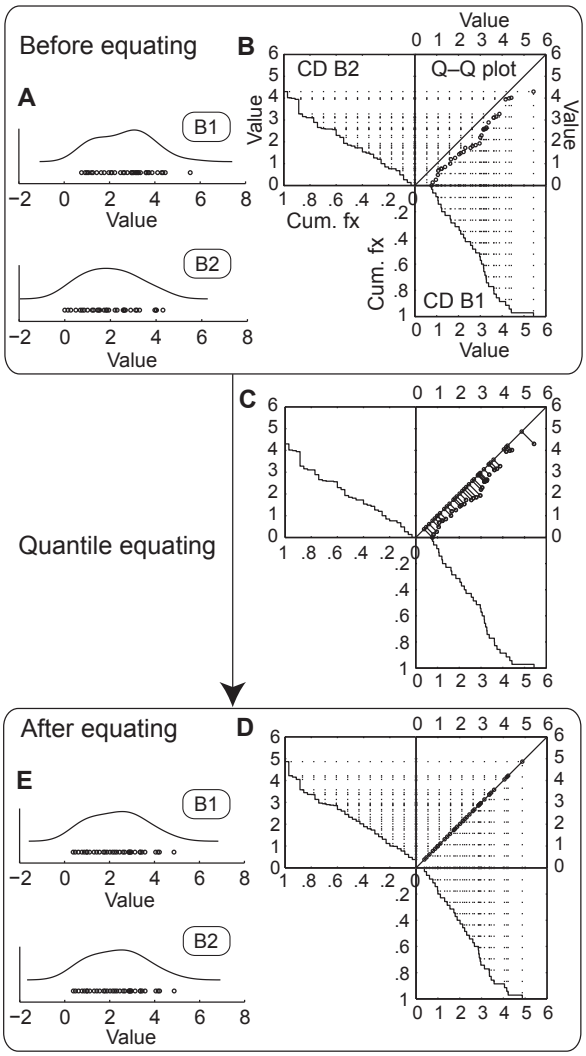

Figure 3.1: Action of quantile equating algorithm schematically illustrated: Data samples B1 and B2 have different distribution shapes (panel A). The cumulative distributions (CD) corresponding to these distributions are plotted against each other in the quantile–quantile plot (Q–Q plot) in panel B. Quantile equating is attained by projecting the values of corresponding quantiles onto a scalar multiple of the unit vector (the diagonal line in the Q–Q plot) in panel C. Then, the projected (averaged) quantile values are substituted for the original values in the subsets belonging to each quantile. Thereby, the distributions of B1 and B2 become equal, as is illustrated with equal cumulative distributions (panel D) and equal kernel densities (panel E). Data from ref[146]. CD, cumulative distribution; Q–Q plot, quantile–quantile plot; Cum. fx, cumulative fraction. The axis labels as in panel B apply to panels C and D as well.

possible cause is that for each block a separate target table was constructed on basis of the QC sample measurements in that block. This might have led to different detection thresholds for the same peaks in both blocks and thereby to systematic differences in peak integrals. The scores based on the B1 and on the B2 plasma NMR measurements overlapped only partially (Figure 3.2C). This may have been caused, at least in part, by different CPMG parameter sets in both blocks. Furthermore, in Figure 3.2C, it can be observed that the NMR measurements in B2 of the 12 individual samples that were measured in B1 as well are not representative for the measurements in B2. We suspect that this is among others due to the additional freeze-thaw cycle that these repeatedly measured samples underwent and that is known to affect plasma NMR spectra.[147] Therefore, Figure 3.2C gives a visual illustration of a case where methods that employ such repeatedly measured samples for equating, *e.g.*, the method described in ref[128], cannot be used.

B1–B2 correlation of quantile values

The average Pearson correlation for all variables between the B1 and the B2 quantile values before equating was 0.97 (SD, 0.03) for the LC–MS data and 0.92 (SD, 0.09) for the plasma NMR data. In case of the LC–MS data, notably a group of TGs displayed nonlinear relationships between the quantile values of both blocks (Supporting Information Table 3.4). Among the lipids, TGs are particularly likely to display nonlinear differences in data distribution shapes among data blocks because they can form dimers during ionization and MS detection. This effect is dependent on concentration and on ion source tuning.

Unlike LC–MS systems, NMR spectrometers are regarded to be linear detectors,[148] implying that signal intensity should be linearly related to compound concentration over the complete dynamic range. Therefore, in case of the NMR data, nonlinear relationships between the distributions of the B1 and the B2 data at lower intensities (Supporting Information Table 3.5) might have been caused by differences in the sensitivity of the NMR probe heads used for the acquisitions of the NMR data between both blocks, as well as by differences in peak detection thresholds between both blocks.

Multivariate parameters

The values of parameters that characterize the similarity of the B1 and B2 data sets in the PC space before and after quantile equating are given in Table 3.1. For both the LC–MS data and the plasma NMR data, the values for the P parameter as well as the values for the C parameter with inclusion of two PCs suggest that the structures of the B1 and B2 data are already comparable before equating (Table 3.1, sections A and C). This is important because it suggests that the compositions of the object groups are indeed balanced between both measurement blocks. Therefore it might be reasonable to assume that with application of the quantile equating method, relatively much analytical

between-block variation will be removed with respect to biological variation.

However, the zero values for the R parameter in case of both the LC–MS as well as the NMR data suggest that there is a multiplicative difference between the B1 and B2 data, which is in concordance with what can be observed in the PCA scores plots on the combined data sets (Figure 3.2, panels A and C). Moreover, in Table 3.1, sections A and C, the values for the C parameter decrease considerably with inclusion of more than two PCs, suggesting that the higher PCs are influenced by differences in data distribution shapes between B1 and B2.

3.4.3 B1–B2 comparison after equating

PCA scores plots

After quantile equating of the data, the systematic nonbiological differences between the B1 and B2 data are not manifest anymore in the PCA scores plots (Figure 3.2, panels B and D). In these plots, the scores based on the individual study samples measured in B1 and B2 are dispersed among each other. Also, the scores based on the measurements of the pooled QC samples in both B1 and B2 are located in the centers of the plots. This is consistent with the expectation that the B1 and B2 pooled QC samples should represent the average sample measured in each of the blocks. Given that this expectation is correct, the location in the centers of the plots of the QC sample measurement scores from both B1 and B2 in turn is a direct consequence of making the data distributions of each variable equal for both blocks by quantile equating.

Multivariate parameters

For both LC–MS and NMR, the increase in the values of the R parameter after equating (Table 3.1 sections B and D) suggests that in particular the distance between the centroids of the B1 and B2 data sets has decreased. The values for the P and C parameters have increased as well. The values for all parameters are not equal to 1 after equating, which is consistent with the notion that although our univariate equating method causes equal or nearly equal data distributions among data blocks at the univariate level, the ranking of objects at this univariate level is retained. Therefore, differences among data blocks at the multivariate level are not necessarily removed by univariate quantile equating as well.

3.5 Conclusions

Combination of semiquantitative metabolomics data sets originating from different measurement blocks where the same metabolites have been measured can be challenging due to nonbiological systematic differences among the blocks. These differences are caused by unwanted, though sometimes practically un-

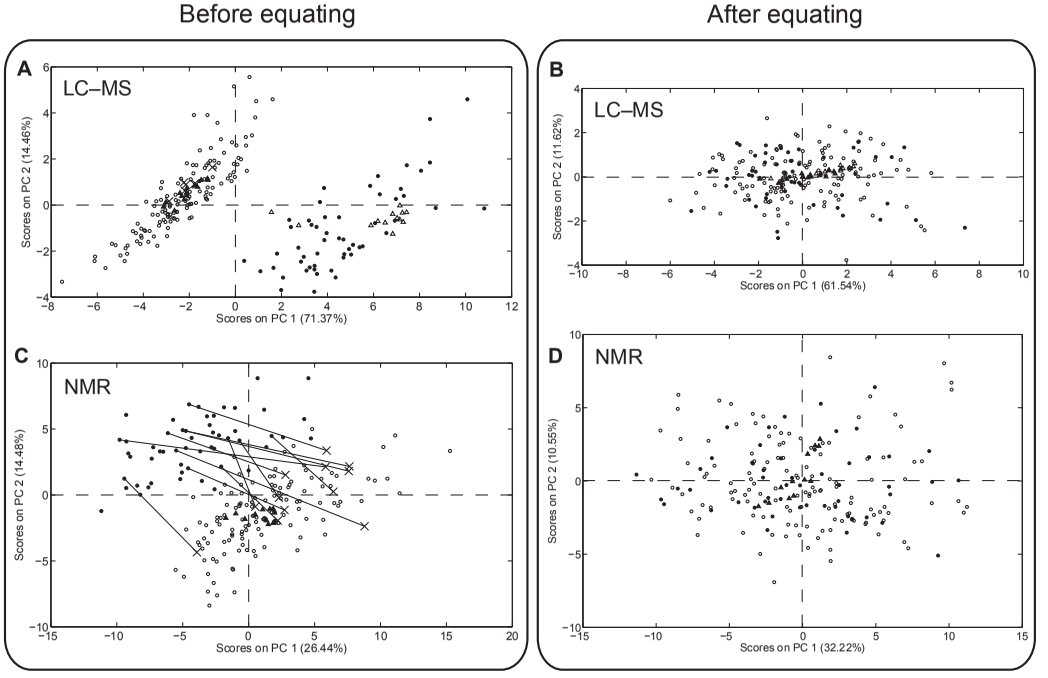

Figure 3.2: PCA scores on PC1 and PC2 for the combined (concatenated) B1–B2 data sets before (panels A and C) and after (panels B and D) quantile equating. Panels A and B, plasma LC–MS data; panels C and D, plasma NMR data. In panels A and B, B1 QC sample aliquots measured in B1 are indicated by (△). In panel C, scores based on NMR measurements of individual plasma samples that were measured in both B1 and B2 are connected by lines. The percentages of variance explained by the respective PCs are given between brackets in the axes labels. PC1–PC2 loadings plots are given in the Supporting Information (Section 3.7 Figures 3.6 and 3.7). ●, B1 individual study sample; ○, B2 individual study sample; ▲, B2 QC sample aliquot measured in B2; ×, B1 QC sample aliquot (panel A) or B1 individual study sample (panel C) measured in B2.

Table 3.1: B1–B2 similarity of data sets in PC space before and after quantile equating[a]

A (LC–MS data, before equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs
P	0.9615	0.9423	0.9339	0.9315	0.9463	0.9513
C	0.9829	0.9504	0.6682	0.6527	0.6181	0.4553
R	0	0	0	0	0	0

B (LC–MS data, after equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs
P	0.9958	0.9952	0.9897	0.9926	0.9954	0.9941
C	0.9984	0.9935	0.9902	0.9844	0.9645	0.9392
R	0.9997	0.9985	0.9988	0.9988	0.999	0.9988

C (^{1}H NMR data, before equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	7 PCs
P	0.949	0.9143	0.9125	0.9057	0.8919	0.8962	0.8936
C	0.9964	0.9947	0.713	0.6732	0.5372	0.3266	0.2944
R	0	0	0	0	0	0	0

D (^{1}H NMR data, after equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	7 PCs
P	0.9892	0.951	0.975	0.97	0.9684	0.9684	0.9679
C	0.999	0.9716	0.805	0.721	0.6402	0.5964	0.5572
R	0.9996	0.9985	0.9866	0.9857	0.9874	0.9879	0.9881

[a]Sections A and B, similarity of B1 and B2 plasma LC–MS data sets before (section A) and after (section B) quantile equating. Sections C and D, similarity of B1 and B2 plasma ^{1}H NMR data sets before (section C) and after (section D) quantile equating. *P*, B1–B2 similarity of PCA loadings patterns; *C*, B1–B2 similarity of variance-covariance matrices; *R*, B1–B2 similarity of data set centroid locations.

avoidable, between-block differences in experimental conditions. We have presented a solution for such data combination problems in the form of the quantile equating method. We have demonstrated the successful application of the quantile equating method to LC–MS and ^{1}H NMR metabolomics data obtained in human plasma samples. We successfully applied our equating method to urine ^{1}H NMR metabolomics data as well (see the Supporting Information for methods and results).

It is conceivable that the quantile equating method is equally applicable for other types of semiquantitative metabolomics data, *e.g.*, GC–MS data. Due to its univariate nature, this equating method will remain to provide satisfactory results even when the data sets to be combined contain data for (much) larger numbers of variables than the examples considered in this article. Moreover, the applicability of the equating method presented in this article may not be limited to data from metabolomics studies. For example, in DNA methylation measurements in the context of epigenetics studies the data distributions may vary between arrays and equating methods have the potential to correct the data obtained in such experiments.

Of course, the possibility to apply equating methods in an "omics" context leaves unimpeded the importance of good analytical practice. This includes that, if possible, all study samples should be measured in one block to minimize process variability. However, in a typical large metabolomics study, where in total hundreds or thousands of samples are measured, it is often not feasible both from a practical and cost perspective to measure new and previously measured samples together in one block. Because of such practical limitations, and because not all systematic differences between measurements in different analytical blocks can be prevented by good analytical practice alone, we believe that equating methods have the potential to enable joint analysis of valuable data sets, which would not be possible without using such methods.

3.6 Acknowledgments

We thank all the twins and siblings who participated in this study. We acknowledge support from The Netherlands Bioinformatics Centre (NBIC) through its research programme BioRange (project no. SP 3.3.1), Spinoza-premie NWO/SPI 56-464-14192, the Center for Medical Systems Biology (CMSB), Twin-family database for behavior genetics and genomics studies (NWO-MaGW 480-04-004), and NWO-MaGW Vervangingsstudie (NWO no. 400-05-717).

3.7 Supporting information

3.7.1 Materials and methods (urine ^{1}H NMR)

Participant recruitment and characterization as well as urine sampling were performed according to the methods described in Chapter 2. In B1 and B2, urine ^{1}H NMR spectra were obtained of nearly all participants of whom blood plasma samples were analyzed as well with LC–MS and ^{1}H NMR (see Sections 3.3–3.4). However, in B1 analysis of the urine sample of one participant was unsuccessful. In B2 the urine sample of one other participant was not analyzed. Without these two participants, the total number of participants of whom urine samples were analyzed in B1 and B2 together was equal to 180. The average ages of the twins of whom urine samples were analyzed in B1 and in B2, and of the siblings, were not different from those of the twins and siblings of whom blood plasma samples were analyzed with LC–MS and ^{1}H NMR. Of four participants only two out of three replicate NMR analyses were successful. In B2, for the purpose of quality control of the NMR analyses QC samples were prepared prior to sample preparation by pooling equal amounts of urine sample from the study participants who were measured in that block.

Before NMR spectroscopic analysis, 1 mL urine samples from all subjects were lyophilized and reconstituted in 700 μL deuterated sodium phosphate buffer (0.1 mmol/L, pH 7.4 made up with D_2O), to minimize spectral variance arising from differences in urinary pH. Sodium trimethylsilyl-[2,2,3,3,-2H4]-1-propionate (TMSP; 0.025 mmol/L) was added as an internal standard for chemical shift. 600 μL of the samples was transferred to 5 mm outer diameter NMR tubes.

Then, the measurement order of the urine samples of the individual study participants was randomized. In B2, after this randomization, uniformly distributed pooled QC sample aliquots were inserted. Furthermore, in B2 following each of these QC sample aliquots, samples were inserted of in total eleven participants that had already been analyzed in B1. These samples thus underwent an additional freeze-thaw cycle between B1 and B2.

NMR spectra were acquired in triplicate on a fully automated Bruker Avance 600 MHz spectrometer (Bruker Analytik GmbH, Karlsruhe, Germany) using a standard 1D ^{1}H NMR pulse sequence with water suppression (zgpr) and operating at an internal probe temperature of 300K. Typically 128 transients were acquired into 64×10^3 data points using a spectral width of 12 kHz; 45 ° pulses were used with an acquisition time of 2.7 s and a relaxation delay of 2 s. The signal of the residual water was removed by a presaturation technique in which the water peak was irradiated with a constant frequency during the relaxation delay.

The spectra were processed using XWIN-NMR software (v.3.1, Bruker Analytik GmbH). The FIDs were multiplied by an exponential weighing function corresponding to a line broadening of 0.3 Hz prior to Fourier transform. The acquired NMR spectra were manually phased, baseline-corrected and referenced

to the TMSP resonance at 0.0 ppm.

The urine NMR data were processed further in the way as described for the plasma NMR data in Section 3.3. Where applicable, names of chemical compounds were assigned to chemical shifts (ppm values) on basis of an in-house reference database.

3.7.2 Results and discussion (urine ^{1}H NMR)

After application of the "80%-rule", 199 features (variables) were kept for further analysis. Typical examples of ^{1}H NMR spectra of urine samples from B1 and from B2 are presented in Figure 3.3.

The consecutive replicate analyses of each sample displayed a decrease of the signal at 4.06 ppm, particularly in B1. Presumably this is a result of progressive exchange over time of methylene protons with deuterium in the creatinine molecule.[149,150] Because this exchange occurred exclusively at the methylene group of the creatinine molecule its effect was observed only in the signal at 4.06 ppm and not in the other creatinine signals in the spectrum. The replicate measurements of the eleven prepared samples that were measured in both B1 and B2 displayed a notable decrease of the signal at this position in B1 but not in B2, presumably because in B2 the exchange had attained a chemical balance situation.

After exclusion of the variable corresponding to this signal from the data, specifically the variables corresponding to the signals at 3.28 ppm and at 3.05 ppm caused separation of the measurements of both years along the first two PCs in PCA (not shown). Presumably this was due to the signals at these chemical shift values to exceed plateau values in the peak detection software in a number of measurements. Prior to median normalization of the data, these two variables were excluded for further analysis as well.

Figure 3.4 shows the results of PCA on the urine ^{1}H NMR data from B1 and B2 prior to between-block effect correction. The scores plot (Figure 3.4A) suggests that there is a multiplicative difference between the B1 and the B2 data, although this between-block effect is not as profound as was seen in case of the plasma LC–MS and NMR data (Section 3.4 Figure 3.2 panel A and panel C, respectively). Compared to these other types of data, in the urine NMR data the within-block variance is relatively larger with respect to the between-block variance. This is probably due to the large biological interindividual variation that is typically observed in urine ^{1}H NMR spectra.

The correlation between the B1 and the B2 quantiles was on average 0.92 (SD 0.08). The variables in the urine NMR data that displayed the lowest Pearson correlations between the B1 and the B2 quantiles are listed in Table 3.2. The values of the parameters that evaluate similarity of datasets in the multivariate space before and after equating are given for the urine NMR data in Table 3.3. The values in Table 3.3A for the P and C parameters suggest that the PCA loadings patterns and variance-covariance matrices are rather similar for the B1 and B2 data even without equating. However, the R parameter val-

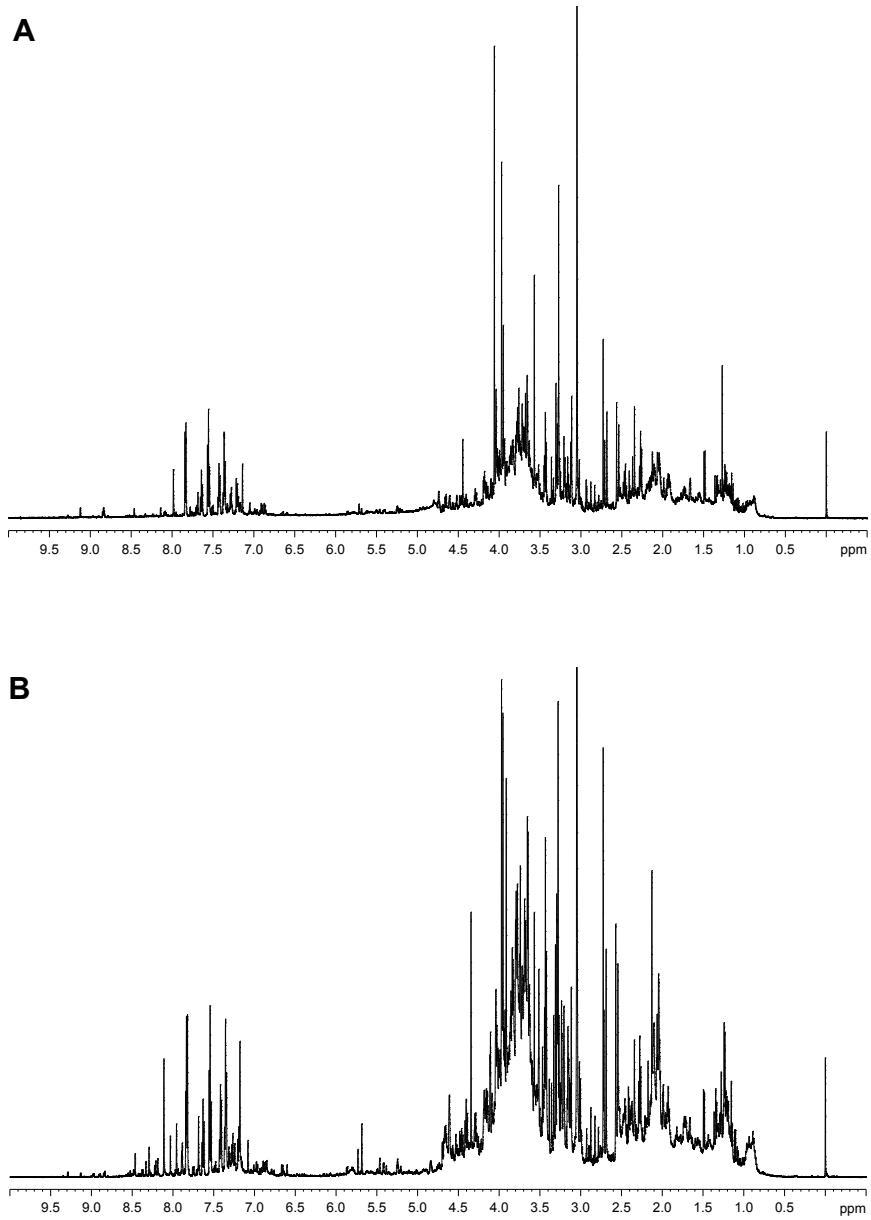

Figure 3.3: Typical ^{1}H NMR spectra of urine from B1 (panel A) and B2 (panel B). Spectra in panel A, and in Figure 3.8A are from the same participant. Similarly, spectra in panel B, and in Figure 3.8B are from the same participant. The signal at 0 ppm originates from the reference standard TMSP.

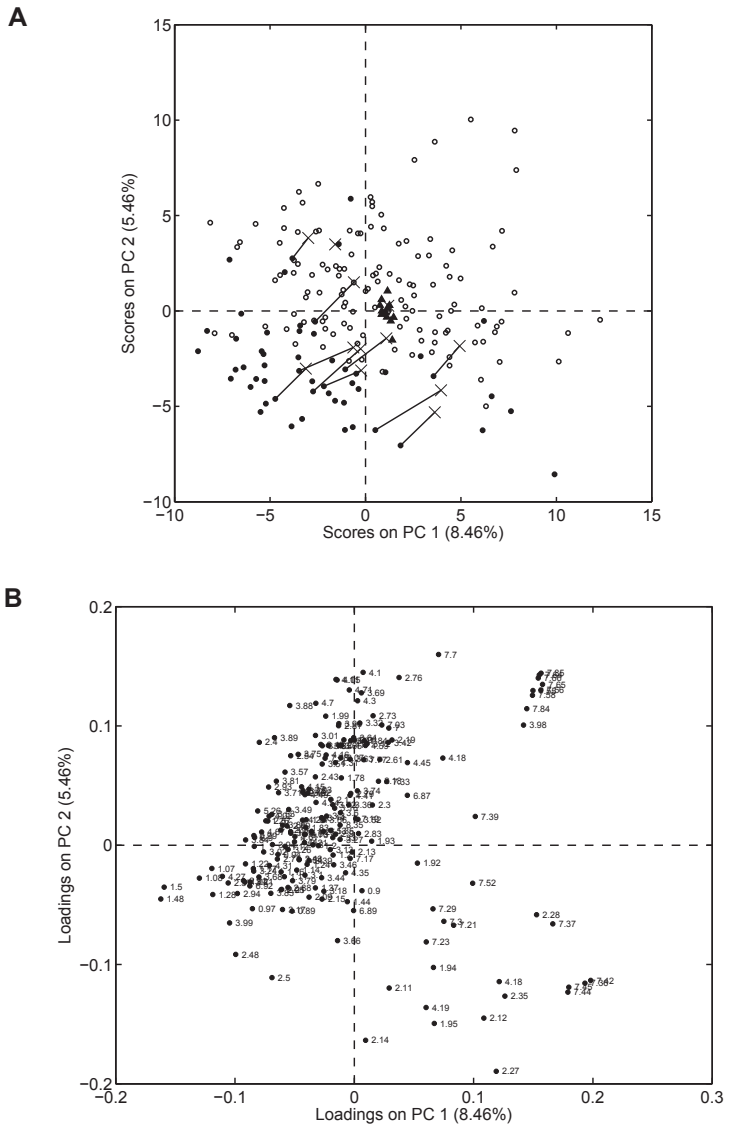

Figure 3.4: PCA scores (panel A) and loadings (panel B) on PC1 and PC2 for the combined (concatenated) B1–B2 urine NMR datasets before correction for between-block effects. Scores based on measurements in B1 and in B2 of individual samples that were measured in both years, are connected by lines in panel A. The percentages of variance explained by the respective PCs are given between brackets in the axes labels. Denotation of markers in panel A: ●, B1 individual study sample; ○, B2 individual study sample; ▲, B2 QC sample aliquot measured in B2; ×, B1 individual study sample measured again in B2. In panel B, loadings are labeled by chemical shift (ppm value).

Table 3.2: Urine ^{1}H NMR features with lowest B1–B2 correlation of quantile values before equating

Chemical shift (ppm)	Pearson's R
1.3343	0.4698
1.3460	0.5386
1.9172	0.6211
4.5407	0.6851
2.1910	0.6883
2.8306	0.7223
8.3517	0.7231
1.2275	0.7302
7.2996	0.7419
3.7926	0.7522

Table 3.3: Similarity of B1 and B2 urine NMR datasets in the PC space before (panel A) and after (panel B) quantile equating [a]

A

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	7 PCs	8 PCs	9 PCs	10 PCs	11 PCs
P	0.9421	0.7716	0.7009	0.7054	0.6991	0.7083	0.6896	0.6773	0.6832	0.6794	0.6794
C	0.9979	0.9538	0.9177	0.8713	0.7964	0.7324	0.6216	0.5367	0.5129	0.4791	0.4302
R	0.9776	0.0566	0.2497	0	0	0	0	0	0	0	0

B

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	7 PCs	8 PCs	9 PCs	10 PCs	11 PCs
P	0.9562	0.8943	0.8821	0.8836	0.876	0.8434	0.8686	0.8759	0.8758	0.874	0.8713
C	0.9947	0.9832	0.9541	0.8945	0.868	0.8158	0.6944	0.5836	0.519	0.4828	0.4312
R	0.9997	0.9969	0.9937	0.9948	0.9956	0.993	0.9931	0.9936	0.9941	0.9944	0.9948

[a] Similarity of B1 and B2 urine ^{1}H NMR datasets before (section A) and after (section B) quantile equating. P, B1–B2 similarity of PCA loadings patterns; C, B1–B2 similarity of variance-covariance matrices; R, B1–B2 similarity of dataset centroid locations.

ues when computed for more than one PC suggest that the centroid locations of both datasets are different prior to equating. This can be seen in Figure 3.4A as well, where the scores of the B1 and of the B2 data are separated mainly along PC2.

Figure 3.5 shows the PCA scores and loadings plots of the B1 and B2 data together after equating. As expected on basis of the relatively small between-block effect as suggested by the PCA scores plot before equating (Figure 3.4A), the scores and loadings plots before (Figure 3.4) and after (Figure 3.5) equating are rather similar as well. Similarly as in case of the plasma LC–MS and NMR data (see Section 3.4 Figure 3.2 panels C and D), the scores based on the measurements of individual samples in B1 and B2 are dispersed among each other after equating (Figure 3.5A). Also, the scores based on measurements of pooled QC sample in B2 are again located in the center of the PCA scores plot. After equating, the patterns of PCA scores of replicate measurements with respect to each other within each block were similar to those before equating (not shown).

The values after equating of the parameters that evaluate similarity of datasets in the multivariate space are given in Table 3.3B. The values in Table 3.3B for the R parameter suggest that the centroid locations of the B1 and B2 urine NMR data with inclusion of more than 1 PC have become much more similar. This is as expected on basis of the nature of the quantile equating method, and can also be observed in Figure 3.5A. The values for the P and for the C parameters have increased slightly as well.

3.7.3 PCA loadings plots for plasma LC-MS and for plasma NMR datasets

Plasma LC-MS

PC1–PC2 loadings plots for the combined (concatenated) B1 and B2 plasma LC–MS datasets before and after equating are given in Figure 3.6.

Plasma NMR

PC1–PC2 loadings plots for the combined (concatenated) B1 and B2 plasma NMR datasets before and after equating are given in Figure 3.7.

3.7.4 Examples of plasma NMR spectra

Typical examples of NMR spectra of plasma samples from B1 and B2 are presented in Figure 3.8.

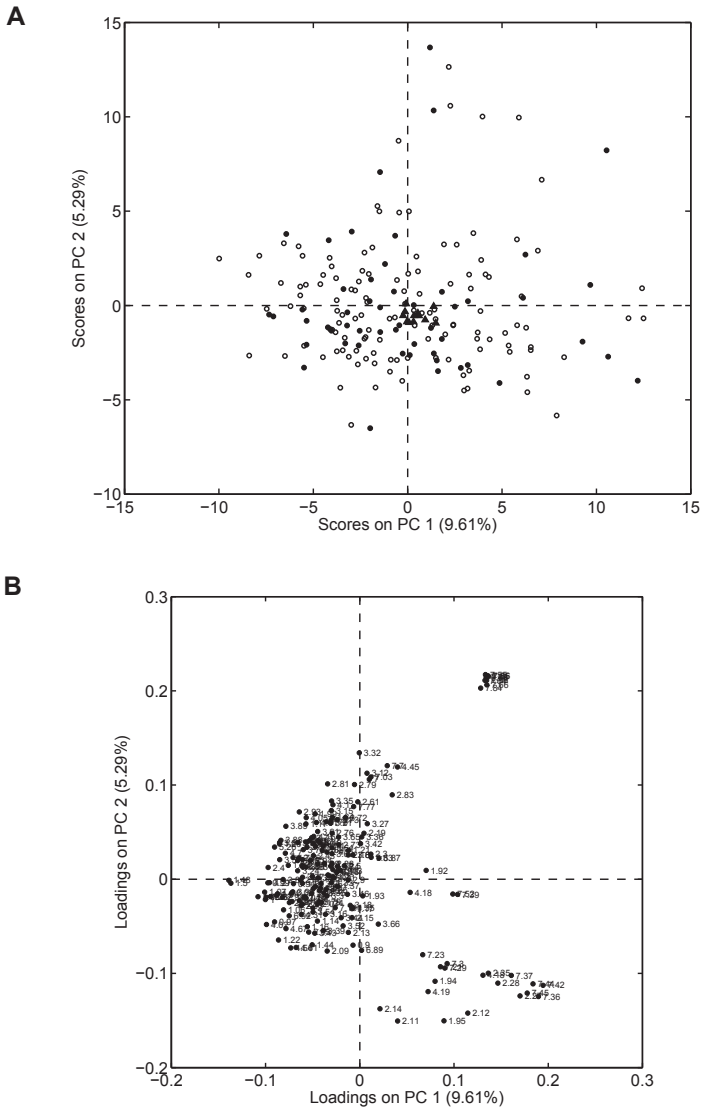

Figure 3.5: PCA scores (panel A) and loadings (panel B) on PC1 and PC2 for the combined (concatenated) B1–B2 urine NMR datasets after quantile equating. The percentages of variance explained by the respective PCs are given between brackets in the axes labels. Denotation of markers in panel A: ●, B1 individual study sample; ○, B2 individual study sample; ▲, B2 QC sample aliquot measured in B2. In panel B, loadings are labeled by chemical shift (ppm value).

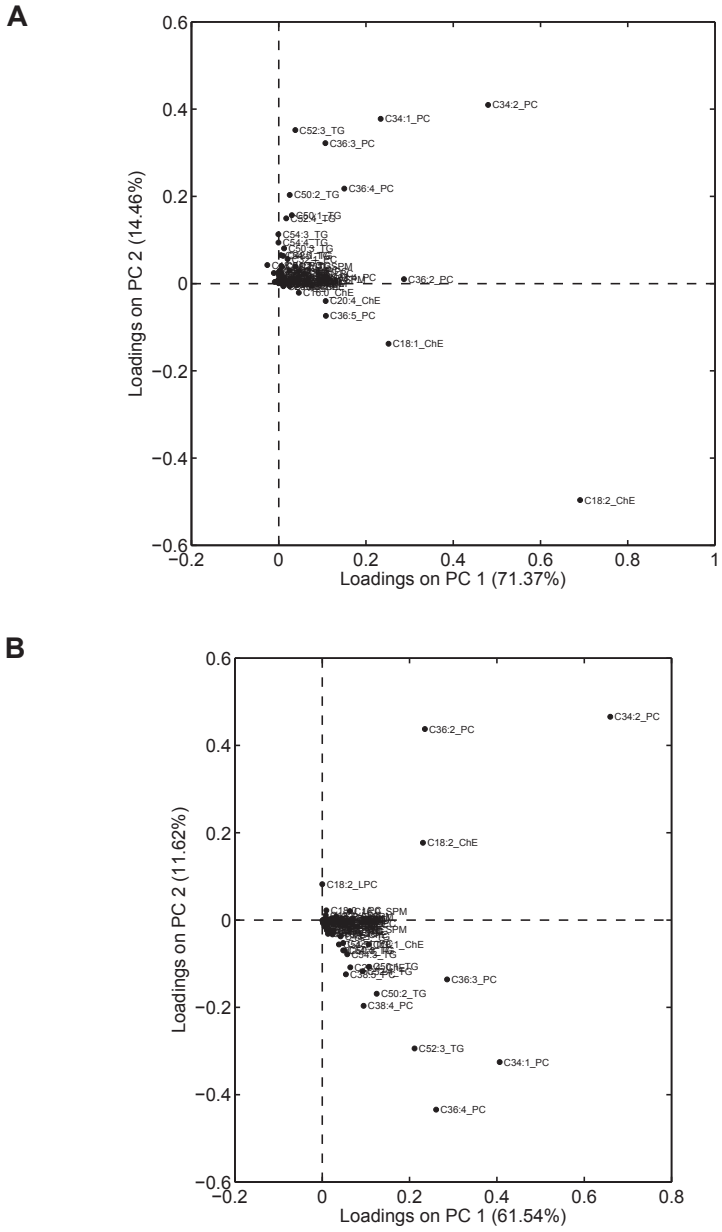

Figure 3.6: PC1–PC2 loadings plots for the combined (concatenated) B1–B2 plasma LC–MS data before (panel A) and after (panel B) quantile equating. The percentages of variance explained by the respective PCs are given between brackets in the axes labels. See Section 3.4 Figure 3.2 panels A and B for the corresponding scores plots.

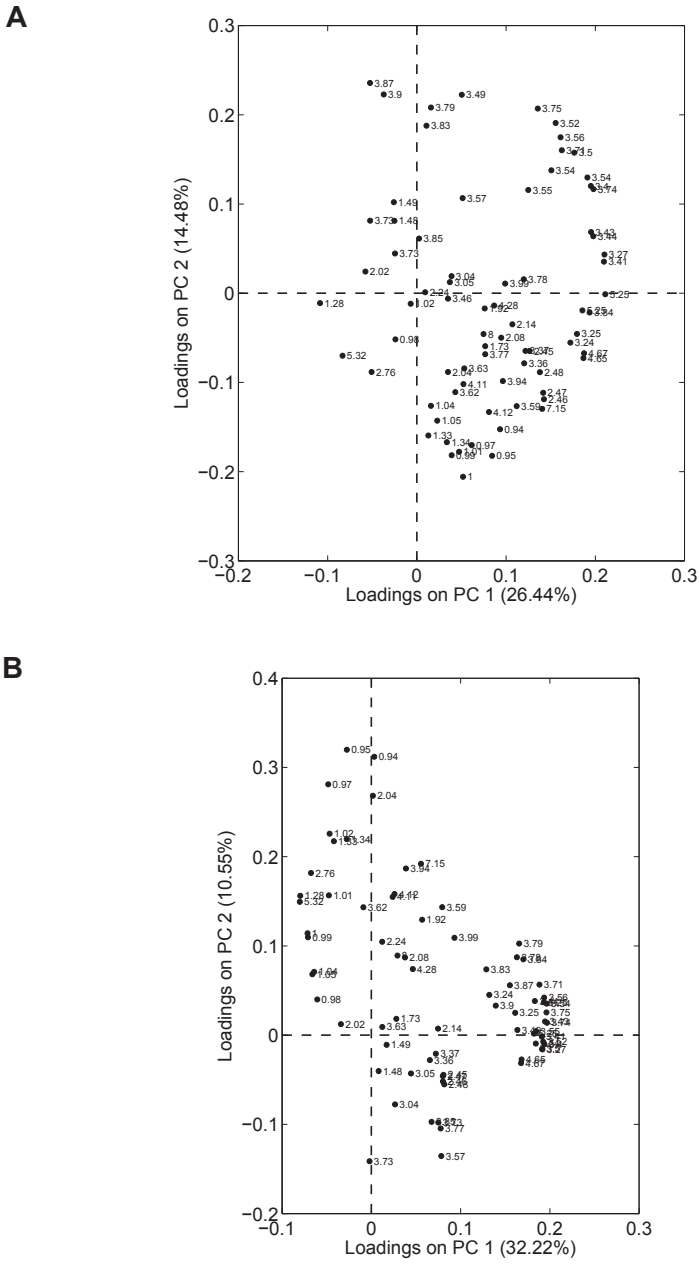

Figure 3.7: PC1–PC2 loadings plots for the combined (concatenated) B1–B2 plasma NMR datasets before (panel A) and after (panel B) quantile equating. The percentages of variance explained by the respective PCs are given between brackets in the axes labels. Loadings are labeled by chemical shift (ppm value). See Figure 3.2 panels C and D in Section 3.4 for the corresponding scores plots.

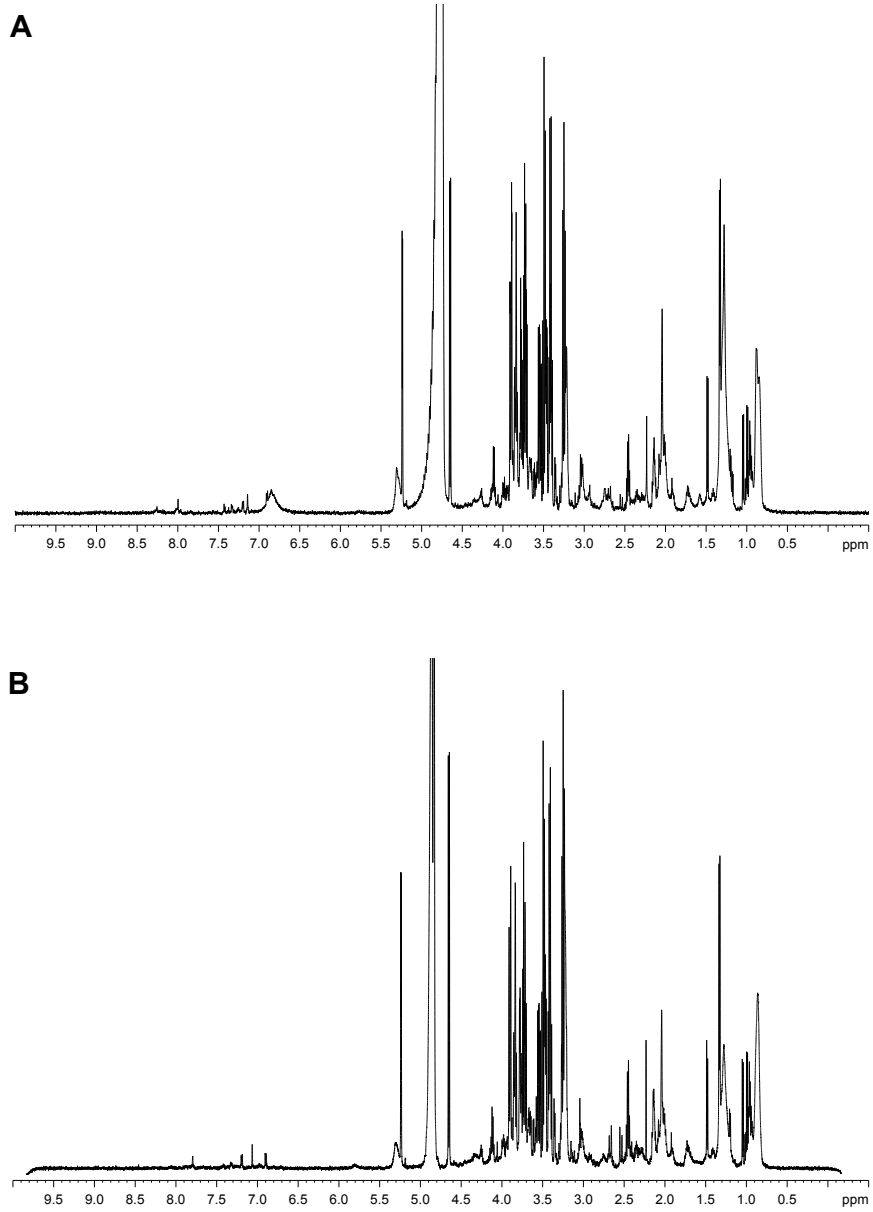

Figure 3.8: Typical ^{1}H NMR spectra of plasma from B1 (panel A) and B2 (panel B). Spectra in panel A, and in Figure 3.3A are from the same participant. Similarly, spectra in panel B, and in Figure 3.3B are from the same participant. Differences between 4.8 and 5.0 ppm in panel A and B are due to the differences in the effectiveness of the water suppression during acquisition of the spectra.

Lipid	Pearson's R
C36:5_PC	0.8254
C48:2_TG	0.8946
C50:4_TG	0.9037
C54:2_TG	0.9106
C48:3_TG	0.9209
C48:1_TG	0.9220
C50:2_TG	0.9283
C50:3_TG	0.9372
C50:1_TG	0.9501
C46:0_TG	0.9521

Table 3.4: Plasma lipids with lowest B1–B2 correlation of quantile values before correction for between–block effects

Chemical shift (ppm)	Pearson's R
1.9238	0.5996
2.0825	0.6546
3.8660	0.7045
0.9753	0.7095
3.7258	0.7409
3.6189	0.7529
1.0087	0.7726
3.9011	0.7737
3.5504	0.7955
3.6256	0.7962

Table 3.5: Plasma NMR features with lowest B1–B2 correlation of quantile values before correction for between–block effects

3.7.5 Variables in plasma LC–MS and NMR data with lowest B1–B2 correlation of quantile values

Plasma LC–MS

The variables in the plasma LC–MS and NMR data having the lowest B1–B2 correlation of quantile values before quantile equating are listed in Tables 3.4 and 3.5.

CHAPTER 4

Hierarchical Clustering Analysis of Blood Plasma Lipidomics Profiles from Mono- and Dizygotic Twin Families

Harmen H.M. Draisma,[1] Theo H. Reijmers,[1] Jacqueline J. Meulman,[2] Dorret I. Boomsma,[3] Jan van der Greef,[1] and Thomas Hankemeier[1]

In preparation for publication

[1]Leiden University, LACDR, Leiden, The Netherlands.
[2]Leiden University, Mathematical Institute, Leiden, The Netherlands.
[3]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands.

4.1 Abstract

Twin and family studies are typically used for the elucidation of the relative contributions of genetic variation and environmental variation to phenotypic variation among individuals. Hierarchical clustering analysis generates an overview of the relative similarities and differences among participants from different families on the basis of multivariate data obtained from these participants. In this study we performed hierarchical clustering analysis on the basis of blood plasma lipidomics data obtained in a healthy cohort consisting of 37 monozygotic twin pairs, 28 dizygotic twin pairs, and 52 of their biological nontwin siblings. These data originated from two separate data sets obtained in different measurement "blocks". In hierarchical clustering analysis of the combined data from both blocks, clustering of the participants in both blocks was dependent on measurement block rather than on family structure. However, after correction of the data for "between-block effects", such clustering of participants according to measurement block was not apparent anymore whereas clustering of family members was still observed. The results of further analyses on the combined, corrected data sets suggested that relative similarities were largest between monozygotic co-twins. The relative similarities between dizygotic co-twins, among sex-matched nontwin siblings and among sex-matched nonfamilial participants were progressively smaller. Dissimilarity of lipid profiles between monozygotic co-twins correlated both with increased levels of the inflammatory marker C-reactive protein and with female gender and, when interpreting the results for males and females separately, with recent illness. Therefore, our results support the hypothesis that shared genetic background and shared environment contribute to similarities in lipidomics profiles. Also, blood plasma lipid profiling appears to be useful for detection and monitoring of disease in individuals. The enhancement of the biological interpretation of data analysis results after correction for "between-block effects" illustrates the beneficial effect of this procedure.

4.2 Introduction

Genetic variation and variation in environmental influences among individuals contribute to individual differences in measurable characteristics, *i.e.* to phenotypic variation. The estimation of the relative contribution of genetic and environmental variation to phenotypic variation is often a first step in the elucidation of the specific causes of individual differences. For such analyses of the heritability[37,151] of traits, often (twin) family studies are used because they are genetically informative and participants within families are relatively well-matched for environmental noise. With respect to heritability analyses using regular families, studies on the basis of twin families[152] have an even enhanced power to detect genetic influences on phenotypic variation.[31] One cause for this is that the members of twin pairs are particularly well-matched

for environmental variation. A second cause is that two types of twin pairs exist, *i.e.* monozygotic (MZ) twin pairs and dizygotic (DZ) twin pairs. MZ twins share all their additive genetic variance whereas DZ twins share only approximately half of their variance at the DNA sequence level; the same degree of additive genetic variance is shared among nontwin siblings.[38] Because of the large difference in shared genetic variance between MZ and DZ twins and the matching for environmental variation between co-twins of both types of twin pairs, comparison of the phenotypic correlations between MZ and DZ twin pairs provides a means to estimate heritability. Such quantitative genetic analyses are often carried out by structural equation modeling (SEM),[38] which provides a univariate estimate for the heritability of a trait.

Quantitative genetic analysis can be performed either for directly outward measurable phenotypes such as height or body weight, or on the basis of measurements of so-called endophenotypes or intermediate phenotypes[10–12] that are physiologically in between the genome and the phenotype. Examples of endophenotypes are gene expression in cells, or levels of proteins or metabolites as measured in body fluids such as blood or urine. Studies of endophenotypes are potentially more informative of the biological pathways leading to the observed phenotypic variation among individuals than the analysis of such phenotypes themselves. Among the endophenotypes, metabolite levels are particularly interesting because metabolites are relatively close to the phenotype and therefore potentially directly relevant for phenotypic variation. Because of their relatively unbiased, comprehensive nature, metabolomics studies capitalize on this because such studies allow for the discovery of novel biological pathways.

When multivariate phenotypic data such as metabolomics data have been obtained in (twin) families, hierarchical clustering analysis (HCA) can be used as an alternative to quantitative genetic analysis on the basis of SEM to obtain an impression of the importance of genetic variation for phenotypic variation. The aim of HCA is to group (*i.e.*, to cluster) objects (for example, family members) such that objects that are relatively similar will be in the same cluster and objects that are relatively dissimilar will be in different clusters.[42] Information regarding group membership is not used during the clustering process; rather, objects that have similar scores on corresponding variables will cluster. The input for HCA is a distance or dissimilarity matrix that represents the dissimilarities among objects on the basis of the multivariate data obtained for each object; the result is a dendrogram (a tree) that represents the relative similarities and differences among objects as a twodimensional structure. When performing HCA of multivariate data obtained in different families, because of the genetic and environmental variance shared by family members it is expected that members of the same family will cluster together and that members of different families will be in different clusters.

A useful property of HCA in general is that it is not hampered by nonpositive definiteness of the input data matrix, and that therefore it is suitable for the analysis of typical "omics" data such as metabolomics data. In the con-

text of (twin) family studies, an advantage of HCA is that it acknowledges the pleiotropic effects of genes influencing the variance of different traits belonging to the same biological pathway. Furthermore, because HCA is an exploratory data analysis technique, in contrast to SEM it allows for the discovery of novel biological effects causing heterogeneity among study participants. As an example of the latter, in Chapter 2 we demonstrated that in HCA of blood plasma lipidomics data obtained in 21 MZ twin pairs, two DZ twin pairs and eight biological nontwin siblings, male and female study participants were separated at the highest level in the clustering dendrogram. This suggested that variance in lipidomics profiles is relatively small among individuals of the same gender.

In this chapter, we report the results of HCA of blood plasma lipidomics data from a healthy cohort of 37 MZ twin pairs, 28 DZ twin pairs, and in total 52 of their biological nontwin siblings. Lipidomics, or the analysis of lipids with metabolomics techniques, is an important part of metabolomics research because lipids are involved in a plethora of (patho)physiological processes.[153] For the current study we combined the data that provided the basis for Chapter 2 with additional data mainly from DZ twin pairs and from biological nontwin siblings. Because these data were measured in different measurement "blocks", we applied the method of "quantile equating" to make the data combinable (see Chapter 3).

The inclusion in the current study of more DZ twin pairs and more nontwin siblings, allowed us to validate and extend our previous observations that have been described in Chapter 2. Also, in this chapter we show that application of quantile equating to make combinable data sets indeed causes biological effects to be visible in the combined data set, rather than non-biological differences between the data from different measurement blocks.

4.3 Materials and methods

4.3.1 Participants

Twins and biological nontwin siblings were recruited from the Netherlands Twin Register.[154] Characterization of participants, collection of fasting blood and urine samples, and sample preparation were performed as described previously.[155–157] Participants completed a number of questionnaires; for the current study, we used answers to questions regarding current use of any medication, recent subjective health, current and earlier smoking habits, and whether participants currently lived at their parents' home. Female participants reported the day of their menstrual cycle at the time of sampling. Zygosity was determined for all twin pairs by DNA genotyping.

4.3.2 Measures

Measurement of C-reactive protein (CRP) concentration and lipidomics profiling in blood plasma samples were performed as described in Chapters 2 and 3.

In brief, lipidomics profiling was performed using an LC–MS method targeted at the analysis of lipids. These measurements were carried out in two "blocks", denoted as B1 and B2, respectively. The measurements of B2 were performed almost one year after those of B1 (see Chapter 3); samples from members of the same family were always measured in the same block. In B1 and B2, one and two replicate measurements were performed per study sample, respectively.

The nonbiological systematic differences between the normalized data from the two measurement blocks were removed by "quantile equating" as described in Chapter 3; the B1 replicate measurements were averaged per study sample prior to equating.

4.3.3 Hierarchical clustering analysis

Clustering analysis of lipidomics profiles was performed using the combined (concatenated with the variables as the shared mode) B1–B2 data sets both before and after application of the quantile equating method, using the methods as described in Chapter 2. That is, first autoscaling was applied to the columns (variables) of the data matrix consisting of the internal standard-corrected responses for all detected lipids in all study participants, with the aim to give all variables equal weight for the subsequent HCA. Subsequently the lipidomics profiles were normalized among individuals (rows) by standard normal variate (SNV) normalization.[94] Then, Euclidean distances among the scaled lipid profiles were computed. SNV normalization followed by computation of the squared Euclidean distances among objects is mathematically equivalent to computing (1–) the correlations among unscaled objects (rows).[96] Euclidean distance matrices were subjected to HCA using the average linkage clustering algorithm, which was chosen on basis of the highest Pearson correlation between the original distance matrices and the cophenetic distance matrices. Heatmaps and associated hierarchical clustering dendrograms were generated using the 'heatmap.2' function in the 'gplots' package in the statistical computing environment `R`.[158]

The remaining analyses, as described below, were performed using the combined B1–B2 data set after quantile equating only. The distributions of the Euclidean distances between MZ co-twins, between DZ co-twins, among nontwin siblings, and among nonfamilial participants were characterized using box plots. To assess whether there were statistically significant differences in median Euclidean distance among MZ co-twins, DZ co-twins, sex-matched nontwin siblings, and nonfamilial participants in the combined equated data set, we performed a multiple comparison procedure using Tukey's honestly significant difference criterion on the basis of the result of a nonparametric analysis of the variance within these groups of study participants versus the variance of the group means.[97] A multiple comparison procedure is designed to be conservative when testing for significant differences for more than one pair of groups.[98]

The stability of the hierarchical clustering based on these distances was assessed by a bootstrap analysis (10,000 resamplings) using the 'pvclust' pack-

Table 4.1: Basic description of participants.[a]

	MZM	MZF	DZM	DZF	Nontwin siblings	Total
Number of participants	34	40	20	36	52	182
Average age in years (standard deviation)	18.1 (0.2)	18.1 (0.2)	18.2 (0.2)	18.2 (0.2)	19.3 (4.7)	18.5 (2.5)

[a]MZM, monozygotic male; MZF, monozygotic female; DZM, dizygotic male; DZF, dizygotic female.

age[101] in R. In a bootstrap analysis, the stability of the clustering is assessed upon randomization of the number of occurrences of each variable in the data set, while keeping the size of the data set equal.

Clustering of family members was assessed by 'node analysis' as described in Chapter 2; that is, the distance between MZ co-twins, DZ co-twins, or a pair of nontwin siblings was assessed as the number of nodes or branching points in the dendrogram separating the members of the pair. For each possible number of nodes separating MZ or DZ co-twins or nontwin siblings in the dendrogram, we compared the observed number of co-twin or sibling pairs separated by that number of nodes, with the number of observations that was expected on basis of chance. Chance distributions were created by permutation of the object labels over the leaves of the clustering dendrogram. Such p-values were computed for each of in total 100 sets of permutations, where each set consisted of 10,000 permutations. On the basis of these 100 permutation tests we computed the average p-values as well as the standard deviations of these average p-values. For these comparisons, we used a critical value of 5% to denote statistical significance.

4.4 Results and discussion

4.4.1 Participants

The combined data sets based on the measurements obtained in the two measurement blocks comprised data on 59 lipids detected in the sample from each participant. The participants originated from in total 65 families; 79 participants were male and 103 were female (see Table 4.1). In one monozygotic female (MZF) family and one dizyotic male (DZM) family, a twin pair and two nontwin siblings (in both families, one male and one female nontwin sibling) participated; in all other families, only one nontwin sibling participated. All DZ twin pairs included in the study were same-sex pairs; 33 of the total 52 nontwin siblings were of the same sex as their twin siblings.

4.4.2 Hierarchical clustering analysis

The results of HCA are displayed as dendrograms with an associated heatmap indicating the Euclidean distances between pairs of objects (Figures 4.1 and 4.2).

The Pearson correlations between the original Euclidean distance matrix, and the cophenetic distance matrix based on HCA of the combined B1–B2 data sets were 0.75 and 0.60 before and after equating, respectively.

Before correction for nonbiological differences between the B1 and B2 data, the objects in the combined (concatenated with the variables as the shared mode) B1–B2 data set clustered very strongly according to the block (B1 or B2) in which they had been measured (Figure 4.1). However, after quantile equating, in the clustering based on the combined B1–B2 data sets, objects measured in the two respective blocks were dispersed among each other (Figure 4.2). This was already expected on basis of the principal component analysis scores plots based on the combined equated B1–B2 data sets (see Figure 3.2B in Chapter 3).

In Chapter 2, in HCA on the basis of the single B1 data set, we had observed that objects segregated almost perfectly according to gender at the highest level in the dendrogram. However, in the dendrograms based on the separate B2 data set (not shown) as well as in the combined B1–B2 data sets both before (Figure 4.1) and after (Figure 4.2) equating, we did not observe such strong clustering of male and female participants. Upon comparison of the structures of the B1 and B2 data sets in the principal component (PC) space using multivariate methods, we had already found slight differences both before and after application of the quantile equating method (Table 3.1 in Chapter 3). That is, both before and after equating we found that the similarity of the B1 and B2 covariance matrices decreased from 3 PCs onwards. Perhaps remarkably, this apparently contradicts the lower average relative standard deviations over all lipids (as computed on the basis of measurements of a quality control sample consisting of pooled individual study samples) in B2 with respect to B1 that we reported in that same publication. A cause for this difference in structure between the B1 and B2 data sets could be that in B1, for each sample two replicate measurements were performed, whereas in B2 each sample was measured only once. Therefore, the averaged replicate measurements in B1 might provide higher precision to estimate the true biological effects, than the single replicate measurements as in B2.

The stability of the clustering on the basis of the combined equated B1–B2 data sets, as assessed by a nonparametric bootstrap procedure, was similar to that observed in the separate B1 data before equating (see Figure 4.5 in Section 4.7; cf. Figure 2.2 in Chapter 2).

In accordance with our previous results using the separate B1 data before equating (Figure 2.1 in Chapter 2), in the combined equated B1–B2 data sets the average Euclidean distance appeared to increase when considering MZ co-twins, nontwin siblings, and nonfamilial participants, respectively (see Figure 4.3). Indeed, the differences in median Euclidean distance between several

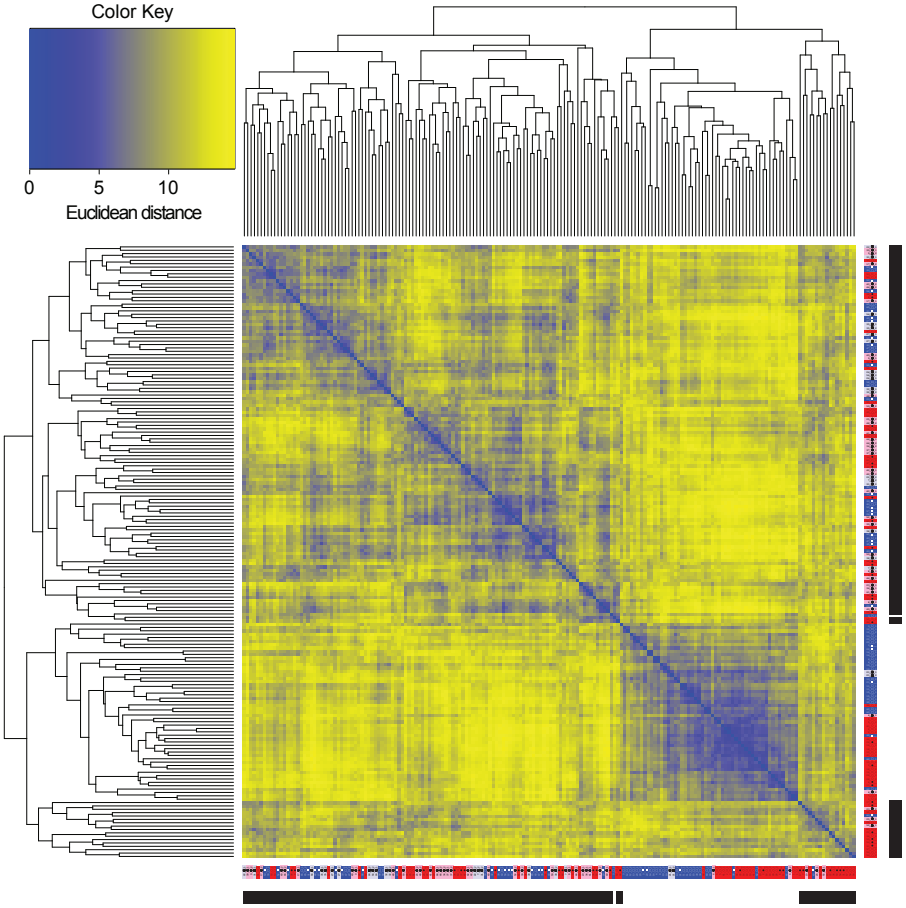

Figure 4.1: Heatmap of Euclidean distances between objects, and associated hierarchical clustering dendrograms for the combined (concatenated with variables as shared mode) B1–B2 data set before quantile equating. In this figure, individual objects are labeled by two color codes: the first color encodes the gender of the participant of whom the sample was obtained (red for females and blue for males). Dizygotic female and dizygotic male twins are indicated with pink and light blue, respectively. The second color encodes the block in which the sample of this participant was measured (white for B1 and black for B2). Participants are denoted as follows: the family identifier (1–65) is followed by a square (□, for males) or a circle (○, for females) to indicate the sex of the participant, and, in case of twins, a "1" or a "2" to indicate the first and second members of the twin pair, respectively. Nontwin siblings are indicated by filled squares (■) or filled circles (●) for males and females, respectively. For the participants from B1, see Table 4.6 in Section 4.7 for a comparison between the labeling as used in Chapter 2 and the labeling used in this chapter.

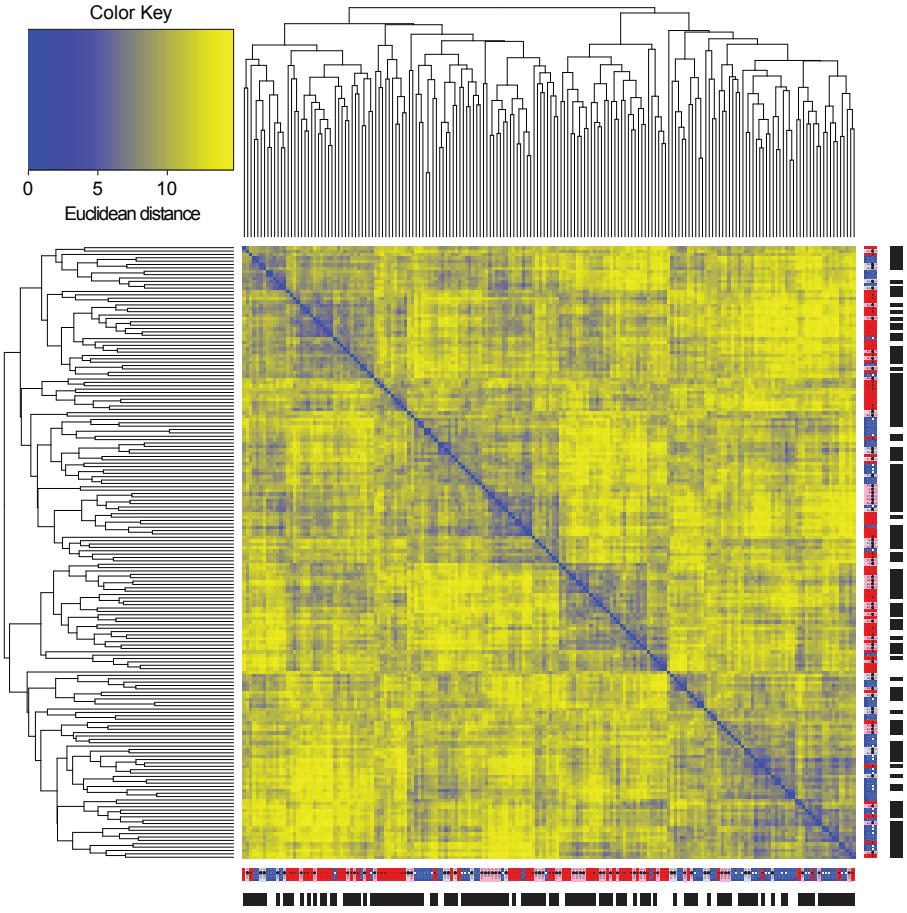

Figure 4.2: Heatmap of Euclidean distances between objects, and associated hierarchical clustering dendrograms for the combined (concatenated with variables as shared mode) B1–B2 data set after quantile equating. For legend, see Figure 4.1.

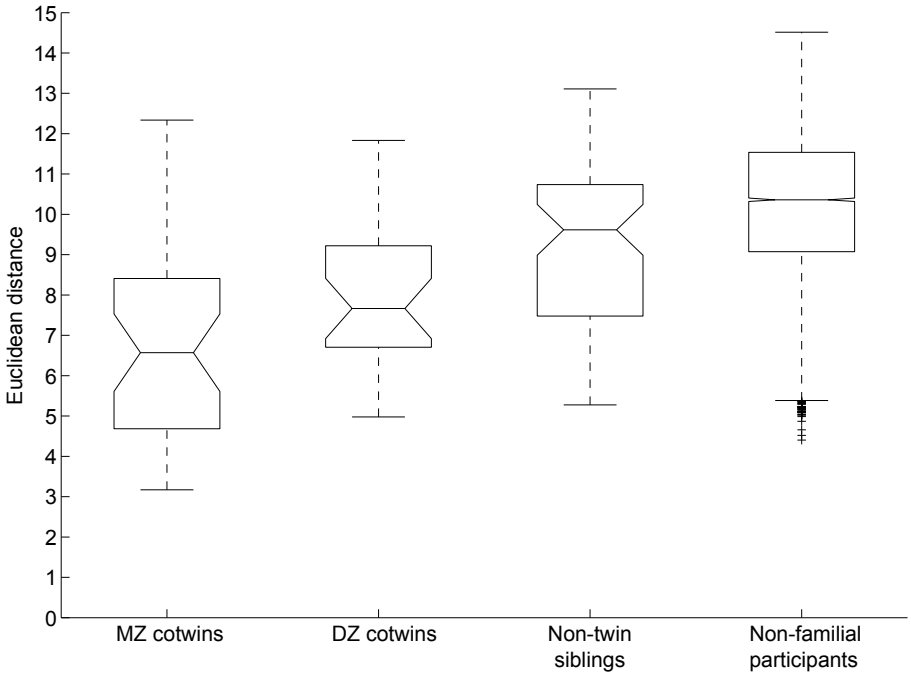

Figure 4.3: Box-whisker plots showing distributions of Euclidean distances between MZ co-twins (N=37), between DZ co-twins (N=28), among sex-matched nontwin siblings (N=66), and among sex-matched nonfamilial participants (N=8,203) in the combined equated B1–B2 data set. The observations indicated with a plus sign in case of the nonfamilial participants illustrate the slight skewness of the distribution of the Euclidean distances among all participants.

Table 4.2: p-values as resulting from multiple comparison test for differences in median Euclidean distances between MZ co-twins, DZ co-twins, sex-matched nontwin siblings, and sex-matched nonfamilial participants [a]

	MZ co-twins	DZ co-twins	Nontwin siblings	Nonfamilial participants
MZ co-twins	-	-	-	-
DZ co-twins	>0.05	-	-	-
Nontwin siblings	<0.01**	>0.05	-	-
Nonfamilial participants	<0.01**	<0.01**	<0.01**	-

[a] **: $p<0.01$

subgroups of participants were statistically significant on the basis of a multiple comparison procedure (see Table 4.2).

Figure 4.3 shows that the average Euclidean distance among biological nontwin siblings assumes a middle ground between the average distance between DZ co-twins and the average distance among nonfamilial participants. This is as expected because, while biological nontwin siblings share on average the same degree of additive genetic variance as do DZ co-twins, the degree of shared environmental variance is less among nontwin siblings than between DZ co-twins.[69]

Clustering of MZ co-twins, of DZ co-twins, and of nontwin siblings in the combined equated B1–B2 data set were characterized using 'node analysis'. The statistical significance of the clustering of family members was assessed by comparison of the observed numbers of occasions where a particular number of nodes separated co-twins or nontwin siblings, with a reference distribution as provided by permutation testing. The results of these comparisons are visualized and summarized in Figure 4.4, and in Table 4.3 in Section 4.7, respectively. In line with our previous results on the basis of the separate B1 data before equating (see Chapter 2), for the MZ twin pairs only the number of occasions (in the current study fifteen) where co-twins were separated by one node in the dendrogram, was significantly larger than the number of occasions that was to be expected on the basis of chance (Figure 4.4A, and Table 4.3A in Section 4.7). However, for the DZ twin pairs, the number of twin pairs separated by one node (four pairs) as well as the numbers of twin pairs separated by five (two pairs), six (three pairs) or nine nodes (three pairs) in the dendrogram were significantly larger than was expected on the basis of the permutation test results (Figure 4.4B, and Table 4.3B in Section 4.7). The relatively small number of DZ twin pairs separated by only one node with respect to the number of MZ pairs separated by one node, as well as the observation that there were also more DZ twin pairs separated by more than one node than was expected on the basis of chance, suggest that the smaller degree of genetic variance shared by DZ co-twins with respect to MZ co-twins contributes to lower relative similarities of DZ twin pairs. This is in concordance with the larger intrapair

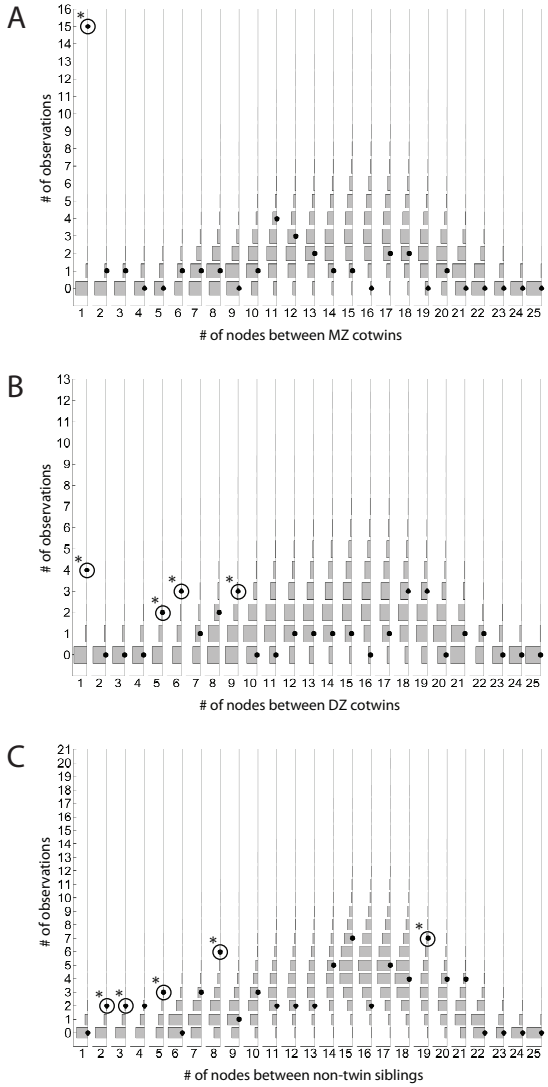

Figure 4.4: Results of node analyses for MZ co-twins (A), DZ co-twins (B), and sex-matched nontwin siblings (C) with respect to permutation-based chance distributions. Numbers of nodes separating co-twins or nontwin siblings increase from left to right in each panel. For each number of branching points, from bottom to top the number of twin or nontwin sibling pairs separated by that particular number of branching points in the permutation tests is displayed by gray bars. Black dots indicate the number of observations given the original ordering of labels along the leaves of the dendrogram as in Figure 4.2, and in Figure 4.5 in Section 4.7. The depicted chance distributions were created by combination of the results from all (*i.e.*, 100) sets of 10,000 permutations. Asterisks indicate average p-values <0.05 (see Table 4.3 in Section 4.7).

Euclidean distances for DZ twins relative to MZ twins.

In the case of the nontwin siblings, we observed no sibling pairs that were separated by one node in the dendrogram, but we did observe significantly larger numbers of pairs than was expected on the basis of the permutation tests that were connected by two nodes (two pairs), or by three (two pairs), five (three pairs), eight (six pairs) or nineteen nodes (seven pairs) (Figure 4.4C, and Table 4.3C in Section 4.7). This might have been due to the fact that the twin pairs included in this study were all approximately 18 years old, whereas the variance in the age of the nontwin siblings was naturally slightly larger (see Table 4.1).

For the nontwin siblings, we used permutation distributions incorporating the fact that in our study based on twin families, each nontwin sibling is always separated from two sex-matched twin siblings. Therefore, in Figure 4.4C, and in Table 4.3C in Section 4.7, the total number of observed frequencies (*i.e.*, 66) is twice as large as the number of sex-matched nontwin siblings in the combined B1–B2 data set (*i.e.*, 33). This is in contrast to the situation for MZ and DZ co-twins, where each twin is separated from only one co-twin.

Nine MZ twin pairs in the combined B1–B2 data of which the co-twins were only separated by one node, came from B1 (these pairs were separated by only one node in the analysis of the separate B1 data as well, see Chapter 2); in this analysis the total number of MZ twin pairs separated by one node was 13. The remaining six pairs of MZ co-twins separated by only one node came from the B2 data. Five of these six pairs were separated by one node in the separate B2 data as well (not shown); in the analysis of the B2 data separately there was one additional pair of MZ co-twins separated by one node. Another pair of MZ twins (belonging to the family with identifier '43', see the legend to Figure 4.1) who were separated by more than one node in the separate B2 data, were separated by only one node in the combined equated B1–B2 data set. This suggests that due to quantile equating, the lipid profiles of the members of this particular MZ pair have been made more similar. This was suggested as well by comparing the dendrograms for the B2 data before and after equating (not shown).

In analysis of both the separate B1 data as well as of the combined equated B1–B2 data set, separation of MZ co-twins by more than one node appeared to correlate with a relatively high average CRP level (see Figure 4.6 in Section 4.7). In this respect, the pair with family identifier "1" (pair "A" in Chapter 2; see also Table 4.6 in Section 4.7) is a remarkable exception: both co-twins have a similar, relatively high CRP level, yet are separated by only one node. This might be explained by the fact that both co-twins had reported recent flu-like symptoms (see Table 4.4 in Section 4.7), perhaps associated with similar changes in lipid profiles.

In Tables 4.4 and 4.5 in Section 4.7, descriptions are given for MZ co-twins separated by only one and by more than one node in the combined equated B1–B2 data sets, respectively. Next to high average CRP, like in analysis of the separate B1 data (see Chapter 2), female gender appeared to correlate

positively with relative dissimilarity of lipid profiles between MZ co-twins. That is, of the 15 MZ twin pairs separated by only one node, only 4 pairs (27%) were female; in contrast, of the 22 MZ pairs separated by more than one node, 14 pairs (64%) were female. Such dissimilarities of lipid profiles between female MZ co-twins might be associated with asynchronous menstrual cycles. Also in accordance with our previous results, when interpreting the results for male and female MZ twin pairs separately it appeared that in general, recent illness correlated positively with separation of co-twins by more than one node.

4.5 Conclusions

In this study, we have extended our previous analyses of the relative similarities of lipidomics profiles between MZ co-twins, DZ co-twins, among biological nontwin siblings, and among nonfamilial participants based on HCA. The statistical power of these analyses was enhanced due to the successful combination of two different metabolomics data sets. In general, the similarities were largest between MZ co-twins; relative similarities between DZ co-twins, among nontwin siblings and among nonfamilial participants were progressively smaller. In concordance with our previous findings on the basis of a cohort consisting mainly of MZ twin pairs, dissimilarity of lipid profiles in MZ twin pairs as assessed by node analysis and permutation testing appeared to correlate positively with relatively high average blood CRP levels and with female gender. The latter correlation might be associated with asynchronous menstrual cycles. Also, within the groups of female and male MZ twin pairs separately, we observed that in general recent illness correlated positively with dissimilarity of lipid profiles between co-twins.

However, in the current study we were unable to replicate our previous finding that in HCA based on the lipidomics profiles of healthy individuals, male and female participants are separated at the highest level in the resulting clustering dendrogram. This might be due to the fact that our previous findings were based on two replicate lipidomics analyses per study sample, whereas of the samples comprising the second data set used in this study only one replicate measurement had been performed.

Taken together, our findings support the notion that shared genetic background and/or shared environmental exposure contribute to similarities in blood plasma lipidomics profiles among individuals. Strong 'environmental' influences such as recent illness appear to accentuate dissimilarities of blood plasma lipids among individuals, suggesting a role for lipid profiling in detection and/or monitoring of disease. Furthermore, the results obtained in this study suggest that the quantile equating technique is useful to make combinable metabolomics data sets, which increases the power of statistical analyses.

4.6 Acknowledgments

We thank all the twins and siblings who participated in this study. We would like to acknowledge support from the Netherlands Bioinformatics Centre (NBIC) through its research programme BioRange (project number: SP 3.3.1); the Netherlands Metabolomics Centre; Spinoza-premie NWO/SPI 56-464-14192; the Center for Medical Systems Biology (CMSB); Twin-family database for behavior genetics and genomics studies (NWO-MaGW 480-04-004) and NWO-MaGW Vervangingsstudie (NWO no. 400-05-717).

4.7 Supporting information

Figure 4.5: Clustering dendrogram on the basis of combined equated B1–B2 data sets, with associated probability values based on nonparametric bootstrap procedure. Numbers near the branching points in the dendrogram indicate bootstrap probability (bp) values; high values indicate high stability of the corresponding node during bootstrapping. The dendrogram structure in this figure is equal to that of the dendrogram displayed at the top of the heatmap in Figure 4.2. For denotation of participants, see the legend to Figure 4.1 in Section 4.4; for the participants from B1, see Table 4.6 for a comparison between the labeling as used in Chapter 2 and the labeling used in this chapter.

Table 4.3: Numbers of MZ co-twins (A), DZ co-twins (B), and nontwin siblings (C) separated by particular numbers of nodes, with respect to chance observation [a]

A

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
II	15	1	1	0	0	1	1	1	0	1	4	3	2	1	1	0	2	2	0	1	0	0	0	0	0
III	0.0*	15.0	20.8	100.0	100.0	46.8	57.6	69.3	100.0	81.4	17.7	43.2	74.8	95.2	97.6	100	87.4	82.0	100.0	85.0	100.0	100.0	100.0	100.0	100.0
IV	0.00	0.30	0.42	0.00	0.00	0.51	0.50	0.49	0.00	0.39	0.34	0.48	0.42	0.20	0.14	0.00	0.34	0.38	0.00	0.40	0.00	0.00	0.00	0.00	0.00

B

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
II	4	0	0	0	2	3	1	2	3	0	0	1	1	1	1	0	1	3	3	0	1	1	0	0	0
III	0.0*	100.0	100.0	100.0	4.4*	1.1*	47.7	21.6	4.3*	100.0	100.0	84.8	87.3	89.9	93.9	100.0	93.6	40.5	26.6	100.0	60.1	44.7	100.0	100.0	100.0
IV	0.00	0.00	0.00	0.00	0.22	0.11	0.48	0.41	0.20	0.00	0.00	0.36	0.28	0.27	0.23	0.00	0.26	0.53	0.40	0.00	0.53	0.50	0.00	0.00	0.00

C

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
II	0	2	2	2	3	0	3	6	1	3	2	2	2	5	7	2	5	4	7	4	4	0	0	0	0
III	100.0	2.4*	4.8*	9.0	2.8*	100.0	12.7	0.7*	80.0	43.7	84.7	87.8	90.5	43.0	25.7	97.2	59.0	66.9	6.0*	28.9	10.0	100.0	100.0	100.0	100.0
IV	0.00	0.14	0.22	0.29	0.16	0.00	0.37	0.09	0.43	0.41	0.35	0.31	0.30	0.49	0.45	0.18	0.54	0.46	0.23	0.52	0.36	0.00	0.00	0.00	0.00

[a] The rows of each panel represent: number of nodes separating co-twins (row I); observed number of occasions where siblings are separated by the number of nodes as given in row I (row II); average p-value ($\times 100\%$) over 100 permutation tests (10,000 iterations per permutation test; direct comparison of the observed frequencies as in row II with the chance distribution generated by each permutation test) (row III); and standard deviation of the p-value ($\times 100\%$) as in row III, over the 100 permutation tests (row IV). Asterisks indicate average p-values <0.05

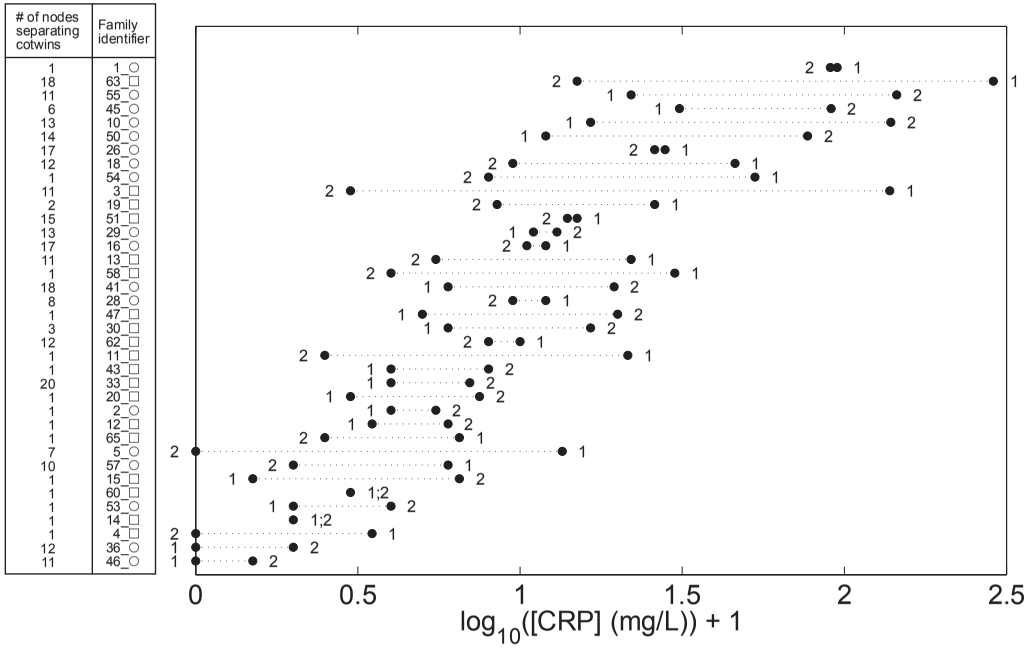

Figure 4.6: C-reactive protein (CRP) levels in blood samples from MZ twins. From bottom to top, the average CRP level of twin pairs increases. On the left hand side, for each MZ twin pair the number of nodes separating co-twins in the dendrogram is displayed. The numbers "1" and "2" near the observations denote the 'first' and 'second' twin of each pair, respectively (for an explanation of the labeling, see the legend to Figure 4.1 in Section 4.4). For producing this figure, the measured CRP concentrations (in mg/L) were $\log_{10}$-transformed to normalize their distribution, and a value of 1 was added to the transformed values to avoid negative numbers. Therefore, for example a value of 2 in this figure corresponds to a measured CRP-concentration of 10 mg/L, which can be considered the upper limit of normal values for CRP concentration in healthy humans.[159] For the participants from B1, see Table 4.6 for a comparison between the labeling as used in Chapter 2 and the labeling used in this chapter.

Table 4.4: Description of MZ twin pairs separated by only one node in the dendrograms of Figure 4.2 and Figure 4.5[d]

Twin pair	*Description*
1_○	Both co-twins had eaten rolls with jam and had drunk soft drink for breakfast at the day of sampling; furthermore, in the sample of 1_○_1 some hemolysis had occurred. Both co-twins had reported recent flu-like symptoms more than one week prior to sampling. This correlated with a rather high average CRP level in this twin pair. Also, the menstrual cycles of both co-twins were not completely synchronous.
54_○	54_○_2 smoked 4 cigarettes per day at the time of sampling while 54_○_1 did not smoke. 54_○_1 and 54_○_2 had had a cold more than one week and more than one month prior to sampling, respectively. This correlated with a relatively high average CRP level in this twin pair.
58_□	Both co-twins used antihistamine as medication for chronic hay fever; 58_□_1 had suffered from hay fever in the week prior to sampling.
47_□	Both co-twins had had a cold more than one month prior to sampling.
11_□	Both co-twins had had a cold more than one month prior to sampling.
43_□	43_□_1 and 43_□_2 had had a cold more than one month and less than one week prior to sampling, respectively. Also, both co-twins had left their parents' home approximately half a year prior to sampling.
20_□	Both co-twins had had a cold more than one month prior to sampling.
2_○	2_○_1 and 2_○_2 had had a cold more than one month and more than one week prior to sampling, respectively. Furthermore, 2_○_2 suffered from allergy.
12_□	12_□_2 had eaten something during the fasting period. Both co-twins smoked at the time of sampling; 12_□_1 had been smoking 15 cigarettes/day for 3.5 years, whereas 12_□_2 had been smoking 8 cigarettes/day for 5 years. Furthermore, 12_□_1 had suffered from fatigue and headache more than one week prior to sampling, whereas 12_□_2 had suffered from flu accompanied by fever more than one month prior to sampling.
65_□	65_□_1 and 65_□_2 smoked 30 and 20 cigarettes/day at the time of sampling, respectively. Both co-twins had smoked less than one hour prior to sampling, and had had a cold less than one week prior to sampling.

[d] For an explanation of the labeling of families and participants, see the legend to Figure 4.1 in Section 4.4.

Table 4.4: Description of MZ twin pairs separated by one node (continued)

Twin pair	*Description*
15_□	Both co-twins had suffered from flu accompanied by fever more than one month prior to sampling.
60_□	Both co-twins had had muesli with diary products for breakfast at the day of sampling. 60_□_1 suffered from chronic back pain and had suffered from stomach flu accompanied by fever more than one month prior to sampling; 60_□_2 had had a cold more than one month prior to sampling.
53_○	53_○_1 and 53_○_2 had suffered from a cold and from stomach ache more than one month prior to sampling, respectively.
14_□	14_□_2 had eaten a roll for breakfast at the day of sampling whereas 14_□_1 had not. 14_□_1 had had a cold more than one week prior to sampling; 14_□_2 had suffered from flu accompanied by fever more than one month prior to sampling.
4_□	4_□_1 and 4_□_2 had had a cold less than one week and more than one month prior to sampling, respectively; furthermore, 4_□_1 suffered from allergy.

Table 4.5: Description of MZ twin pairs separated by more than one node in the dendrograms of Figure 4.2 and Figure 4.5[e]

Twin pair	*Description*
46_○	46_○_1 had reported sickness and headache more than 1 week prior to blood sampling. However, this did not correlate with a high average CRP level for this twin pair. Both twins had synchronous menstrual cycles, although 46_○_2 appeared to suffer from oligomenorrhea.
3_□	3_□_1 had self-reportedly been ill without having a fever less than 1 week prior to blood sampling; this correlated with a high blood plasma CRP level in this participant.
5_○	5_○_1 had smoked in the past (2 cigarettes/day) for half a year 1.5 years prior to blood sampling. Furthermore, 5_○_2 had had a cold less than one week prior to sampling. Also, the co-twins did not have completely synchronous menstrual cycles.
10_○	Both twins had self-reportedly suffered from a cold less than 1 week prior to blood sampling. In the blood plasma of 10_○_2, a high CRP level was measured.
13_□	13_□_1 had had a cold less than 1 week prior to blood sampling; this correlated with a higher CRP level than his co-twin.
62_□	62_□_2 had suffered from infectious mononucleosis more than 1 month prior to sampling; this did not, however, correlate with a relatively high CRP concentration in this twin. Moreover, during sample handling, in the sample of this twin hemolysis had occurred.
16_○	16_○_2 had been smoking five cigarettes per day for 6 years and had smoked 2 h before blood sampling; 16_○_1 had quit smoking a half year prior to sampling, after having smoked 10 cigarettes per day for 5 years. Furthermore, 16_○_2 had had a half cup of sugared tea for breakfast on the day of blood sampling. Both twins did not have synchronous menstrual cycles.
18_○	18_○_1 had self-reportedly suffered from flu-like symptoms less than 1 week prior to blood sampling; this correlated with an increased blood plasma CRP level in this participant. Both twins did not have synchronous menstrual cycles.
28_○	Twin 28_○_2 had been using the drug Fluoxetine for depression. Both twins did not have synchronous menstrual cycles.
30_□	30_□_2 had had a sip of cola during the fasting period prior to sampling. Both co-twins smoked at the time of sampling. 30_□_2 suffered from hay fever.

[e]For an explanation of the labeling of families and participants, see the legend to Figure 4.1 in Section 4.4.

Table 4.5: Description of MZ twin pairs separated by more than one node (continued)

Twin pair	*Description*
41_○	Both twins had self-reportedly been ill less than 1 week prior to blood sampling: 41_○_1 had suffered from a cold, whereas 41_○_2 had had flu-like symptoms accompanied by fever. 41_○_2 used oral contraceptives while 41_○_1 did not; furthermore, their menstrual cycles were not synchronous.
45_○	More than one week prior to sampling 45_○_1 had had a cold. 45_○_2 had suffered from stomach flu more than one week prior to sampling, which correlated with a rather high CRP level.
50_○	In the week prior to sampling, 50_○_2 had suffered from nausea and fatigue whereas 50_○_1 had not. 50_○_1 used terbinafine hydrochloride while 50_○_2 did not.
51_□	More than one month prior to sampling, 51_□_1 had had a cold and 51_□_2 had suffered from flu with fever, respectively.
55_○	Both co-twins did not have synchronous menstrual cycles. Furthermore, both co-twins had had a cold in the week prior to sampling.
57_○	57_○_1 suffered from chronic hay fever; 57_○_2 suffered from chronic asthma, for which she used budesonide/formoterol as medication.
63_□	63_□_1 had suffered from flu with fever and laryngitis in the week prior to sampling, for which she used feneticilline. This correlated with a high CRP level in this participant. Also, in the blood sample from 63_□_1 some hemolysis had occurred. 63_□_2 suffered from irritable bowel syndrome. Furthermore, 63_□_2 had smoked in the past (15 cigarettes/day), and had quit smoking two years prior to blood sampling after having smoked for two years.
26_○	26_○_1 had had a cold more than one week prior to sampling; 26_○_2 suffered from severe eczema for which she used a corticosteroid cream as a medication, from lymphedema in a leg, and from chronic respiratory disease.
29_○	In the blood sample of 29_○_2, hemolysis had occurred; furthermore, 29_○_2 had left her parents home about 4 months prior to sampling, while 29_○_1 had not. Both co-twins had had a cold in the week prior to sampling, and their menstrual cycles were not completely synchronous.
33_□	Both co-twins used fluticasone propionate as medication for slight asthma.
36_○	No tentative explanation for non-coclustering on basis of available information

Table 4.5: Description of MZ twin pairs separated by more than one node (continued)

Twin pair	*Description*
19_□	19_□_1 had been ill and 19_□_2 had had a cold more than one month prior to sampling, respectively

Table 4.6: Conversion table between labeling in this chapter and labeling in Chapter 2 for families from B1[a]

Family label in this chapter	Family label in Chapter 2
1_○	A_○
2_○	B_○
3_□	C_□
4_□	D_□
5_○	E_○
6_○	F_○
10_○	G_○
11_□	H_□
12_□	I_□
13_□	J_□
14_□	K_□
15_□	L_□
16_○	M_○
18_○	N_○
19_□	P_□
20_□	Q_□
21_□	R_□
28_○	S_○
30_□	T_□
41_○	U_○
46_○	V_○
60_□	W_□
62_□	X_□

[a]For an explanation of the labeling of families and participants, see the legend to Figure 4.1 in Section 4.4.

CHAPTER 5

Contribution of Genetic and Environmental Factors to Variation in the Human Blood Plasma Metabolome: a Multivariate Study in Twins and Siblings

Harmen H.M. Draisma,[1] Theo H. Reijmers,[1] Jacqueline J. Meulman,[2] Dorret I. Boomsma,[3] Jan van der Greef,[1] and Thomas Hankemeier[1]

Adapted from Draisma et al., *submitted for publication*

[1]Leiden University, LACDR, Leiden, The Netherlands.
[2]Leiden University, Mathematical Institute, Leiden, The Netherlands.
[3]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands.

5.1 Abstract

Phenotypic data obtained in a genetically informative population sample of individuals can be used for quantitative genetic analyses to elucidate the relative contributions of genetic and environmental variance components to the observed phenotypic variation. Metabolomics aims at the comprehensive measurement in a given sample of all small molecules that are intermediate or end-products of cellular metabolism. Therefore, data as provided by metabolomics experiments represent a snapshot of the physiological state of an organism and are particularly informative of actual phenotypic traits such as disease. By structural equation modeling, we analyzed data obtained with two metabolomics methods in blood plasma samples from in total 163 participants (healthy mono- and dizygotic twins and their sex-matched nontwin siblings). Relative concentrations were obtained of 59 individual lipid metabolites by 'targeted' liquid chromatography–mass spectrometry (LC–MS); a 'global' overview of the relative concentrations of metabolites from different metabolite classes was provided by proton nuclear magnetic resonance (^{1}H NMR) spectroscopy. Univariate quantitative genetic analyses of the LC–MS data revealed potentially biologically relevant differences in heritability for different lipids. In multivariate analysis, we observed that in particular lipids of the same lipid class shared genetic causes of phenotypic variance. In contrast, the heterogeneity of genetic causes of phenotypic variation among different metabolites was relatively large in the ^{1}H NMR data. In conclusion, in this study we have shown the potential of uni- and multivariate quantitative genetic analyses to generate biological insight into the importance of genetic variation for variation observed in human metabolomics data.

5.2 Introduction

Recently, the results of the first genome-wide association (GWA) studies have been reported linking genomic variation and variation in human metabolomics data.[12,51,52] Metabolomics is the comprehensive study of the reagents, intermediate products, or end products of cellular metabolism.[2] Being intermediate phenotypes, with respect to studies on the level of actual phenotypes metabolites provide more insight into biological pathways underlying phenotypic variation.[10–12] The study of (endo)phenotypic variation in quantitative traits, such as metabolite levels measured in body fluids might be of relevance for our understanding of the causes of common diseases.[47,66] Among the various endophenotypes that are measured at "omics" scale (*e.g.*, proteins, gene transcripts), metabolites have the most direct link to cellular physiology and functioning.[8,9,160] The measurement at "omics" scale using the currently available analytical techniques provide an unprecedented scale of resolution, which can be even more directly linked to cellular physiology than is possible on the basis of measurements of 'conventional metabolites'.[161] Metabolomics studies

aim to obtain a comprehensive view of all metabolites from particular metabolite classes (in a so-called "targeted" approach), or of the metabolites from all classes (in a "global" approach). Both approaches allow for the discovery of previously unknown biological pathways on the basis of patterns of relationships among different metabolites, which would be much harder to achieve in a classical reductionist approach that focuses only at select compounds.[162,163]

Here we report the results of quantitative genetic analyses of metabolomics data obtained in a genetically informative sample of individuals, *i.e.* in mono- and dizygotic twin pairs and their nontwin siblings. Instead of elucidating the measureable or 'manifest' genotypic variables (*i.e.*, single-nucleotide polymorphisms indicating quantitative trait loci), as is done in *e.g.* GWA studies,[49,50] in such analyses the causes of phenotypic variation are often modeled as latent variables in a structural equation model.[1,38] Analysis of the covariance structure in phenotypic data by structural equation modeling (SEM)[32] allows for the decomposition of phenotypic (co)variance into variance components attributable to genetic variation and to environmental variation. Using select study designs it is also possible to elucidate the relative contribution of gene-environment interaction to phenotypic (co)variation of traits.[29,38,68]

However, this is not possible on the basis of the classical twin design, which is based upon the comparison of the phenotypic covariances of mono- and dizygotic twins raised together. Monozygotic (MZ) twins, who are fertilized from the same egg, share 100% of their additive genetic variance.[29] Dizygotic (DZ) twins, who are fertilized from two separate egg cells, share only on average 50% of their segregating genes; this percentage is the same for biological nontwin siblings. Therefore, any excess phenotypic correlation between MZ co-twins over that between DZ co-twins is an indication that genetic effects contribute to the variance of a trait.[1]

In SEM, such reasoning is formalized in the structural model and its consistency with the observed data is statistically tested.[32] Analysis of the phenotypic covariances for a single trait of mono- and dizygotic twin pairs raised together allows for the estimation of the heritability of this trait, *i.e.* of the proportion of phenotypic variation attributable to genetic variation among individuals.[37] Next to such a univariate analysis, multivariate analysis is used to elucidate the contribution of genetic and environmental effects to the phenotypic covariance among multiple traits, and it increases statistical power to detect genetic effects.[164,165]

In addition to MZ and DZ twins, sex-matched nontwin siblings of these twins were included in this study because this is known to enhance the power to detect genetic as well as shared environmental effects.[166] In this study we performed both uni- and multivariate quantitative genetic analyses using two types of metabolomics data obtained in blood plasma samples from the same participants. That is, we analyzed data from liquid chromatography–mass spectrometry (LC–MS) of plasma lipids, and from one-dimensional proton nuclear magnetic resonance (^{1}H NMR) spectroscopy. The LC–MS data provide a 'targeted' view of the lipid metabolites present in the samples; lipids are

involved in a number of important (patho-)physiological processes.[153] Proton NMR spectroscopy, on the other hand, aims at a more 'global' view of metabolites from different classes, for example amino acids, lipoproteins and carbohydrates.[2] However, with this latter method one can not discriminate among for example the individual lipid metabolites that are detected by the targeted LC–MS method used in this study. Also, with NMR spectroscopy typically only metabolites present in higher concentrations in a sample can be detected.[2]

In our analyses of the LC–MS data, we observed marked heritability for a number of lipids, but also different degrees of heritability among lipids belonging to different classes. In particular, we found a potentially biologically relevant pattern of heritabilities among the lipids of the triglyceride class. In multivariate analysis, in general lipids of the same class tended to cluster together. This suggests that positive phenotypic correlation among blood plasma lipids from the same lipid class is caused by pleiotropic genes.

Probably due to the "global" nature of the used ^{1}H NMR method, the results of the multivariate analyses of the ^{1}H NMR data suggested a much larger diversity in genetic causes of variance for different metabolites than in case of the LC–MS data.

5.3 Materials and methods

5.3.1 Participants

Twins and biological nontwin siblings were recruited from the Netherlands Twin Register.[154] Collection of fasting blood samples from all participants, and sample preparation were performed as described previously.[155–157] Zygosity was determined for all twin pairs by DNA genotyping.

5.3.2 Measures

Semiquantitative metabolomics analyses of the samples obtained from all study participants were performed in two "blocks", where in the first block samples from different participants were analyzed than in the second block (see Chapter 3). Blood plasma was analyzed both with an LC–MS method targeted at the analysis of lipids, and with ^{1}H NMR spectroscopy, as described in Chapter 3 as well. In a metabolomics context, the term "semiquantitative" indicates that no absolute concentrations were measured for the individual metabolites. Rather, we measured either the concentrations of lipids with respect to those of a limited number of so-called "internal standards" (in case of the LC–MS analyses), or the relative concentrations of metabolites with respect to each other (in case of the ^{1}H NMR analyses).

The measurements of the second 'block' were performed almost one year after those of the first 'block'; samples from members of the same family were always measured in the same block. For the data obtained with both methods, the nonbiological systematic differences between the normalized data from the

two measurement blocks were removed by quantile equating (see Chapter 3). This allowed the combination of data from the same variables measured in both blocks into one common data set that can be analyzed with methods like those used in this chapter. After equating, replicate measurement data were averaged per study sample before entering them into SEM as described below.

In this chapter, individual lipid compounds (*e.g.*, C16:1_LPC) as measured with LC–MS are denoted as follows: the number of carbon atoms (*e.g.*, C16) as well as the number of double bonds (*e.g.*, 1) in the lipid, separated by a colon are followed by the class abbreviation (*e.g.*, "LPC" for lysophosphatidylcholines).[127] Proton NMR variables are denoted by the chemical shift values that correspond to the detected features (see Chapter 3).

5.3.3 Genetic analysis

With respect to the quantitative genetic analyses, in this study we followed a similar strategy as was pursued by Schmitt *et al.* in the analysis of voxel-based magnetic resonance imaging data.[167] That is, first we performed univariate genetic analyses to estimate the proportions of phenotypic variance of each variable separately attributable to genetic and specific environmental variance.

Then, we performed all possible bivariate analyses to estimate the genetic and non-genetic components of covariance between all pairs of variables within each data set. The results of these multiple bivariate analyses populated for each data set a genetic correlation matrix, which was subsequently subjected to hierarchical clustering analysis. Schmitt *et al.*, in their 2008-paper, refer to this methodology as "multistep multivariate analysis". A "multistep multivariate" analysis strategy is actually a workaround that provides "semimultivariate" results in cases where existing covariance-based multivariate data analysis methods can not directly be applied to analyze the data for all variables within a data set simultaneously.[168] Typically, as is the case in maximum likelihood-based SEM, data that consist of a relatively small number of objects (participants) and a very large number of (correlated) measured variables prohibit the straightforward use of such existing methods because variance-covariance matrices computed on the basis of such data are non-positive definite.

Variance components were estimated by SEM approach using full information maximum likelihood (FIML) under normal theory using the raw data as input. FIML allows structural equation models to be fitted in the presence of missing values in the data (*e.g.* on twin pairs without nontwin sibling). For SEM we used the novel package OpenMx (version 0.4.1-1320),[169] which is implemented in the statistical computing environment `R`[158] (version 2.10.1).

Univariate analyses

Before fitting variance component models to the data, we established the likelihood resulting from fitting saturated models, where as many characteristics of the observed data (means, variances, covariances) as possible are freely esti-

mated. Then, we equated means and variances within families, and compared the resulting likelihood with that of fitting the saturated model to the data using a likelihood ratio chi-square test. The significance of variance components was tested in a similar way, *i.e.* by comparing the likelihood of the more complex model with that of a more parsimonious model.

We based our choice for a particular genetic variance components model on the customary rules of fit and parsimony, *i.e.*, overall for most variables the fit of the genetic model had to be non-significantly different from that of a saturated model, and overall for most variables the fit of a more parsimonious model (*e.g.*, "E") had to be significantly different from that of the more complex model (*e.g.*, "AE"). Here, the capitals "A" and "E" denote the latent additive genetic and non-shared environmental sources of phenotypic variance, respectively.[38] p-values lower than 0.05 were considered statistically significant. We chose one variance components model (*i.e.*, the "AE" model) to be used for the analysis of all variables, as the sample size was relatively small in the current study. One consideration for doing so was that the estimated values of variance components are always (slightly) dependent on the particular model used, and therefore to be able to compare variance component estimates among different variables it is important that they all have been estimated under the same model.

The homogeneity of means, variances, and variance components across sexes was assessed by comparing the fit of variance components models fitted to the data for males and females separately, with the fit of models where these parameters had been equated across the sexes. For the analysis of all data sets we used data for nontwin siblings only if they were of the same sex as their twin siblings, because the statistical significance of the estimated variance components was higher than when we included opposite-sex nontwin siblings as well (not shown).

Standardized variance components estimates were obtained by dividing the squared values by total variance.[170] Confidence intervals (CIs) for the standardized genetic variance components were likelihood-based.[171]

Bivariate analyses

The components of covariance for each pair of variables within each data set were estimated by fitting a bivariate model based upon a so-called Cholesky composition of the expected covariance matrix (see Fig. 1.4). For initial analysis a multivariate model based upon Cholesky composition is attractive because it is relatively hypothesis-free.[172] For the bivariate analyses, the relative contributions of the same latent sources of phenotypic variance (*i.e.*, "A" and "E") were estimated as in the univariate analyses.[173]

Genetic correlations were computed from the results of the bivariate analyses as follows:[38]

$$r_{\mathrm{x,y}} = \frac{varA_{\mathrm{xy}}}{\sqrt{(varA_{\mathrm{x}} \times varA_{\mathrm{y}})}} \tag{5.1}$$

where $r_{x,y}$ is the genetic correlation between a pair of variables, $varA_{xy}$ is the unstandardized genetic component of the covariance between the two variables, and $varA_x$ and $varA_y$ are the unstandardized genetic components of variance for the respective variables. For each data set, the genetic correlations for each pair of variables were aggregated into a square genetic 'correlation matrix', of which the dimensions equal the number of variables in the data set.[167]

5.3.4 Hierarchical clustering analysis

We used hierarchical clustering analysis to discover patterns of relationships among different variables in the genetic correlation matrices.[174,175] The aim of hierarchical clustering analysis is to group (cluster) variables on the basis of their relative similarities and differences, such that variables that are relatively similar will be grouped together, and variables that are relatively dissimilar will be in different clusters. For hierarchical clustering, we computed the dissimilarities among variables as (1−correlation).[41,176] Then, we subjected the resulting 'dissimilarity matrix' to hierarchical clustering, using the average linkage clustering algorithm. It has been noted[41] that average linkage in practice often performs satisfactorily. The results of hierarchical clustering were visualized using the "heatmap.2" function from the "gplots" package in R.

Of note, as an alternative to hierarchical clustering analysis, eigenvalue decomposition (spectral decomposition, EVD) of the genetic correlation matrix could be used to visualize the patterns of genetic relationships among different metabolites. The eigenvectors as resulting from EVD of a correlation matrix are equivalent to the "loadings" that would result from a principal component analysis on the autoscaled original two-mode (*i.e.*, objects × variables) data matrix on which the correlation matrix was based.[23] By EVD of the genetic correlation matrix, the genetic covariance among variables (metabolites) can be summarized by projecting the original variables onto new orthogonal variables, the so-called principal components (PCs), on the basis of the dominant direction of the genetic covariance among all metabolites. However, we do not show the results of EVD of the genetic correlation matrix here, because hierarchical clustering analysis was used in the remainder of this thesis to summarize the relationships among either objects or variables.

5.4 Results and discussion

5.4.1 Participants

The combined data sets, based on the measurements obtained in the two measurement blocks, comprised data for in total 130 twins and 33 sex-matched nontwin siblings for both LC–MS and ^{1}H NMR. The LC–MS data set contained data on 59 lipids detected in the sample from each participant. Lipids from the following five classes were detected: lysophosphatidylcholines (LPCs); phosphatidylcholines (PCs); sphingomyelins (SPMs); cholesterol esters (ChEs);

Table 5.1: Basic description of participants [a]

	MZM	MZF	DZM	DZF	Nontwin siblings	Total
Number of participants	34	40	20	36	33	163
Average age in years (standard deviation)	18.1 (0.2)	18.1 (0.2)	18.2 (0.2)	18.2 (0.2)	19.0 (4.7)	18.3 (2.1)

[a]MZM, monozygotic male; MZF, monozygotic female; DZM, dizygotic male; DZF, dizygotic female.

and triglycerides (TGs). The ^{1}H NMR data set contained data on 74 features (peaks) detected in each spectrum.

In total 67 participants were male and 96 were female; participants originated from in total 65 families (see Table 5.1). All DZ twin pairs included in the study were same-sex pairs.

5.4.2 Univariate variance components analyses

Genetic models that incorporated heterogeneity of means, variances and covariances across sexes did not fit differently to the data than models where the values for these parameters had been equated across males and females. Therefore, we estimated covariance components for males and females together.

For all data sets, the likelihood-based CIs were rather large, due to sampling error because of the relatively small number of participants in this study. The univariate results specific for each data set are given below.

LC–MS lipids

The heritability estimates per lipid, as well as the 95%-CI, are shown in Figure 5.1. For all measured lysophosphatidylcholines (LPCs), the estimates for the standardized genetic variance components were rather high (range, [0.64–0.75]). The phosphatidylcholines (PCs), on the other hand, displayed relatively much heterogeneity with respect to their heritability: whereas for some lipids (notably C36:2_PC) the estimated heritability was very low, for others (*e.g.*, C36:4_PC) it was rather high. The total range of the estimated heritabilities for the lipids in this class was [0.25–0.77]. The heritability estimates for the sphingomyelins (SPMs) displayed a similar pattern as those for the LPCs: the estimated values for all lipids in this class were rather high (range, [0.47–0.71]). The cholesterol esters (ChEs) displayed a remarkable heterogeneity in their estimated heritabilities when considering the number of C-atoms in the fatty acid: whereas for the measured ChEs with 16 or 18 C-atoms in the fatty acid the estimates were moderate and in the range [0.42–0.48], for the lipids in this class with 20 or 22 C-atoms in the fatty acid the estimates were notably higher and in the range [0.71–0.74].

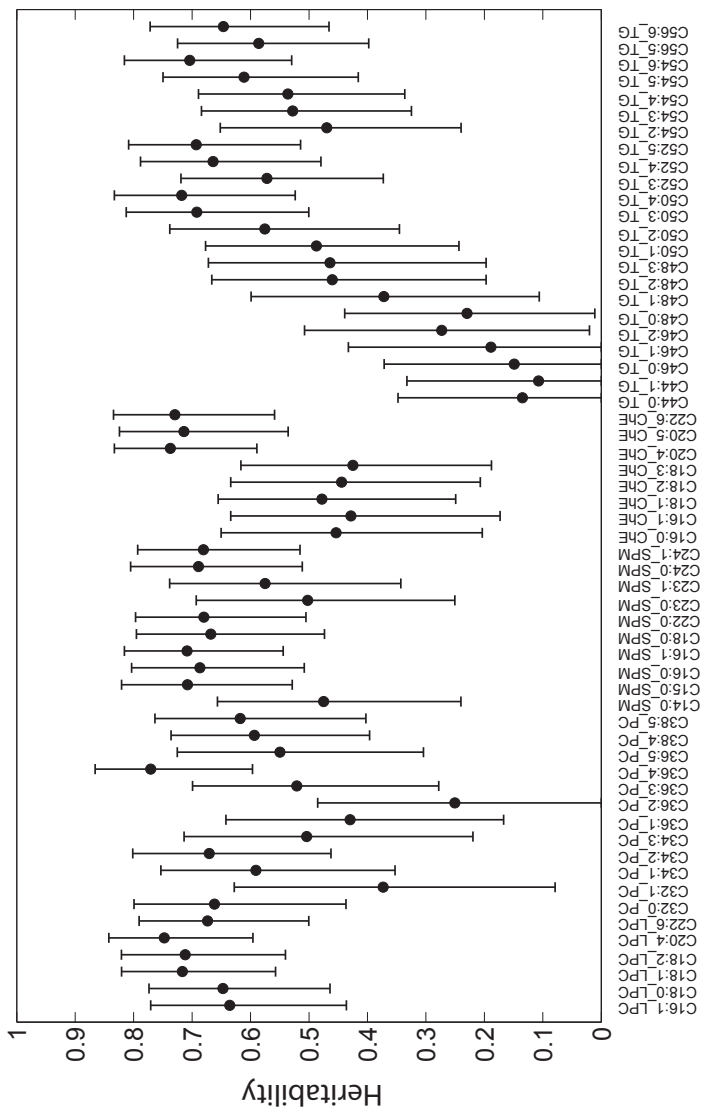

Figure 5.1: LC–MS lipid data: heritability estimates for all lipids under "AE" model. Dots indicate the original estimate for the standardized "$varA$" variance component under a univariate "AE" model; the whiskers denote the maximum likelihood-based 95%-confidence interval for this estimate. Because under the "AE" model, the values of the standardized "$varA$" and "$varE$" variance components add up to a value of one, the values of the "$varE$" variance component estimates can be inferred from this figure as well. For denotation of lipids, see Section 5.3.2.

In the triglycerides (TGs), the pattern was even more striking when considering the combination of the number of C-atoms as well as the number of double bonds in the fatty acids. For 44 up to 50 carbon atoms in the triglyceride, on average the heritabilities increased with additional carbon atoms in the fatty acid. From 50 up to 56 carbon atoms, on average the heritabilities did not change much. However, for each group of TGs with the same number of carbon atoms, with exception of C44, we observed a consistent upward trend in the heritability with increasing numbers of double bonds in the fatty acids. For example, for the TGs with 54 carbon atoms, the estimate for the heritability was always larger for lipids with larger numbers of double bonds in the fatty acids. These remarkable differences in heritability among TGs with different numbers of carbon atoms as well as different numbers of double bounds in the fatty acids might be due to different numbers of conversions by enzymes involved in both catabolism and anabolism of fatty acids. The apparently lower average heritabilities of TGs with numbers of carbon atoms decreasing from 50 up to 44, are perhaps due to increasing numbers of C2-fragment cleavages from the fatty acid backbone (during anabolism) by β-ketoacyl-CoA thiolase, and/or smaller numbers of C2-fragment attachments to the fatty acid backbone (during catabolism) by fatty acid synthase.[177] Similarly, the increases in heritability of TGs with the same number of carbon atoms but increasing numbers of double bonds in the fatty acid backbones, are perhaps due to increasing numbers of actions by enoyl-CoA isomerase and/or 2,4-dienyol-CoA reductase and 3,2-enoyl-CoA isomerase (during fatty acid catabolism), and/or smaller numbers of conversions by fatty acyl-CoA desaturases during fatty acid anabolism. Overall, for the TGs the heritability estimates were in the range [0.11–0.72].

Plasma ^{1}H NMR

The heritability estimates per variable, as well as the 95%-CI, are shown in Figure 5.2. Within the plasma ^{1}H NMR data there was much heterogeneity in the estimated heritabilities among different variables; this is as expected because in contrast to for instance the targeted LC–MS method used to generate the lipid data described in this chapter, NMR is considered a 'global' metabolomics method that should be able to detect metabolites of a much larger number of different classes (*e.g.*, amino acids, carbohydrates). It is conceivable that different classes of metabolites are subject to different (genetic and/or environmental) mechanisms that influence their phenotypic variance. Therefore, in data from a global method like NMR, it is expected that (widely) different relative contributions of genetic and environmental causes of variance are estimated for different metabolites.

Assignment of compound names on basis of the estimated heritabilities of the features detected in the ^{1}H NMR spectra alone is difficult, amongst others because the same compound may have a signal at multiple positions in the spectrum. Also, it is often difficult to elucidate which metabolites corre-

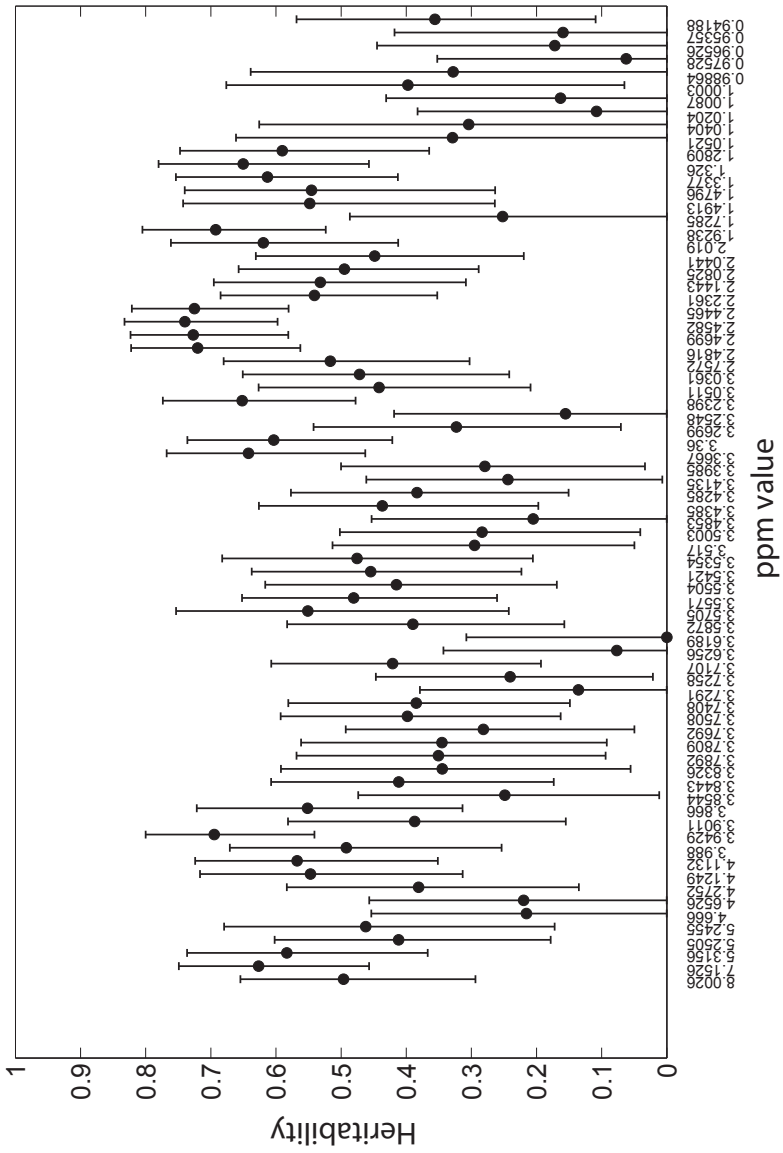

Figure 5.2: ^{1}H NMR data: heritability estimates for all features under "AE" model. For legend, see Figure 5.1. Features (variables) are indicated by their corresponding chemical shift (ppm) value; the variables are sorted from left to right along the horizontal axis in this figure in keeping with the order of the features as these occur in the original NMR spectrum, *i.e.* from high to low.

spond to the measured ppm values because peaks of multiple metabolites may overlap.[178] In our case, most compounds that we putatively linked to a particular combination of features (ppm values) on basis of an in-house reference database, did not display a consistent pattern of heritability for all features within such a combination.

5.4.3 Multivariate analyses

The heritabilities as computed on basis of the bivariate analyses resembled those as resulting from the univariate analyses; this is in line with previous findings.[173] The results of the multivariate analyses specific for each of the two data sets are given below.

LC–MS lipids

Figure 5.3 displays a heatmap of the dissimilarities that result from rescaling the genetic correlations, as well as the associated dendrogram resulting from hierarchical clustering based on these dissimilarities. For most pairs of lipids, the genetic correlations were larger than zero: the median correlation was 0.49 (range, [-0.41; 1]). This suggests that most of the lipids detected in this study have at least some common genetic causes of phenotypic variance. All LPCs clustered together perfectly; the TGs also clustered together very well although one PC (C36:2_PC) clustered together with the TGs because of a very high genetic correlation. For the clustering among the TGs, the number of double bonds in the fatty acid appears to be important: in Figure 5.3, two main clusters of TGs can be observed where one cluster consists of TGs with up to two double bonds in the fatty acid, whereas the TGs in the other cluster have two or more double bonds in their fatty acid chains. This may indicate the action of different enzymes in the metabolism of the TGs in the two different clusters. The SPMs also clustered together rather well, although the results suggest that they share genetic causes of variance with three ChEs (*i.e.*, C16:0_ChE, C18:1_ChE, and C18:2_ChE) as well. The PCs also have a tendency to cluster, although the clustering pattern suggests that also the lipids in this class share some genetic causes of variance with notably the ChEs.

Plasma ^{1}H NMR

Figure 5.4 shows the heatmap of the dissimilarities that result from rescaling the genetic correlations, as well as the associated dendrogram resulting from hierarchical clustering based on these dissimilarities. The median genetic correlation among the ^{1}H NMR variables was 0.08; range [-1; 1]. Note that this is in contrast with the situation for the LC–MS data, where almost all genetic correlations were larger than zero. This contrast might indeed be due to the fact that the LC–MS method used to generate the data analyzed in this study is a 'targeted' method that detects metabolites of the same class (in this case lipids) that indeed may share an important part of their biological pathways

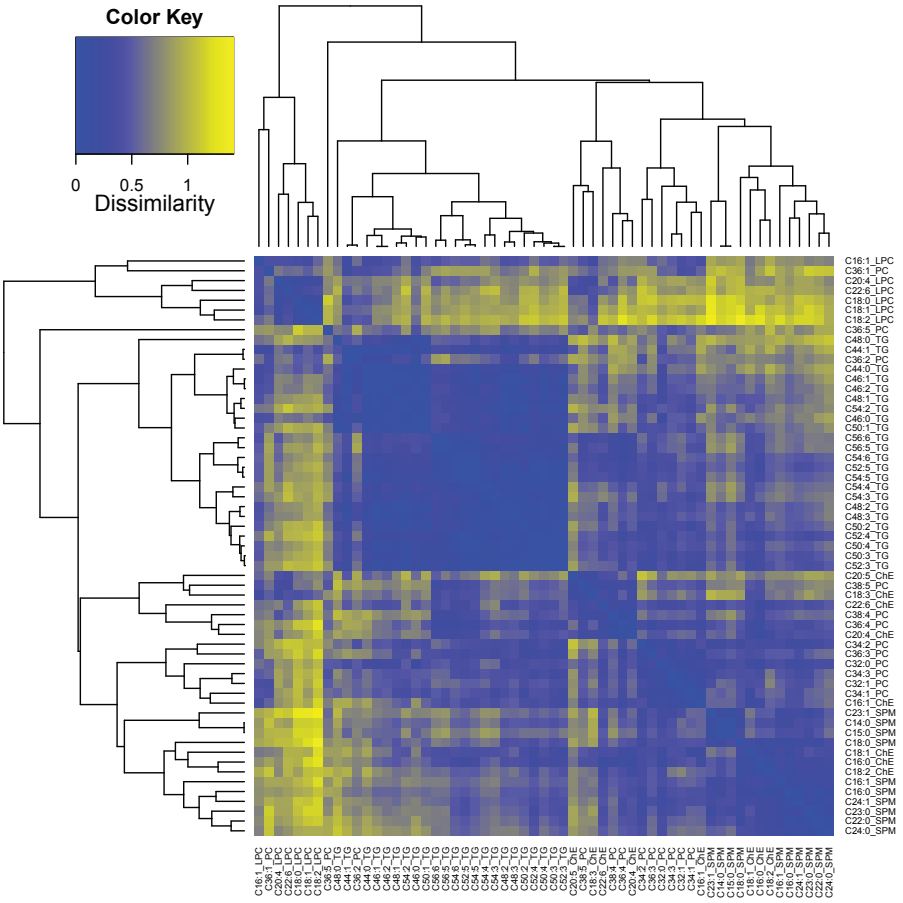

Figure 5.3: LC–MS lipid data: analysis of genetic correlation matrix through dissimilarities. The heatmap indicates with a color code for each pair of lipids the dissimilarity that results from rescaling of the genetic correlation as explained in Section 5.3. For example, a dissimilarity equal to zero as displayed in this figure corresponds to a genetic correlation of 1; a dissimilarity of 1 corresponds to a genetic correlation equal to zero. The average linkage algorithm was used for hierarchical clustering based on these dissimilarities; the resulting dendrogram is shown both along the horizontal and the vertical axes of the ordered heatmap. The Pearson correlation between the cophenetic distance matrix estimated from the dendrogram, and the original dissimilarity matrix based on genetic correlations, was equal to 0.77. The dissimilarity matrix was treated as being symmetric for producing this figure. Therefore, this figure is symmetric with the diagonal of the heatmap as the axis of symmetry; the dendrograms along the horizontal and vertical axes of the heatmap are mirrors of each other. For explanation of lipid labeling, see Section 5.3.2.

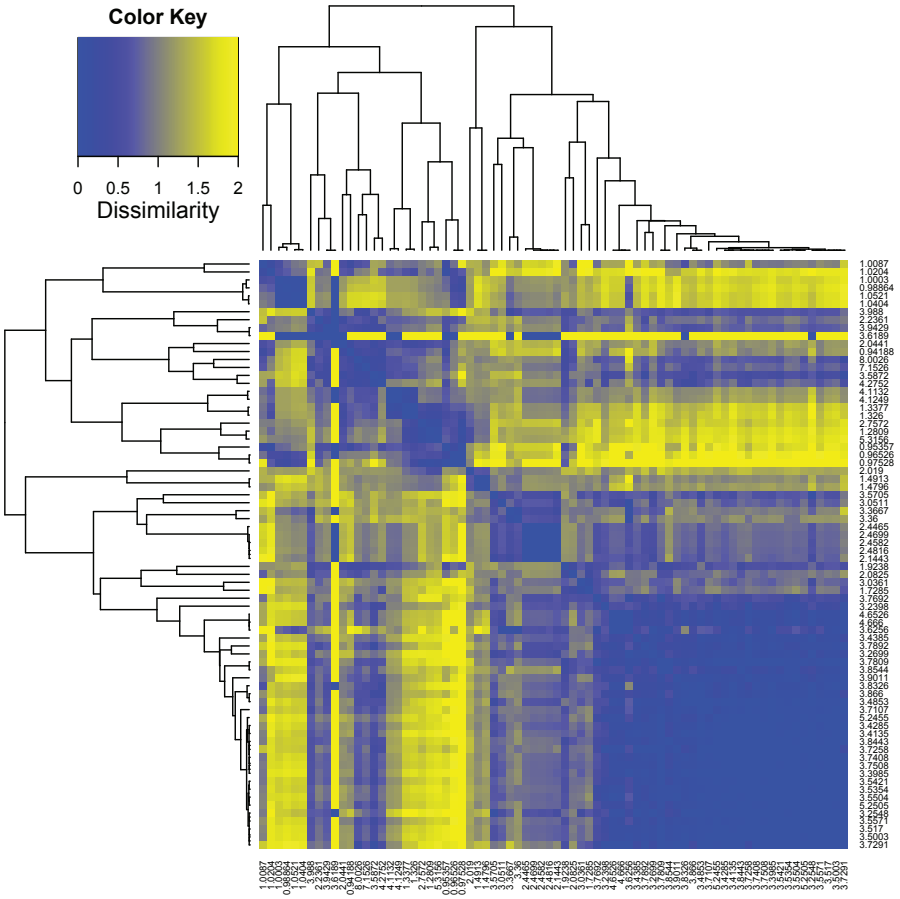

Figure 5.4: ^{1}H NMR data: analysis of genetic correlation matrix through dissimilarities. For explanation, see the legend to Figure 5.3. The Pearson correlation between the cophenetic distance matrix estimated from the dendrogram, and the original dissimilarity matrix based on genetic correlations, was equal to 0.79. Variables are denoted by the chemical shift values that correspond to the detected features.

to phenotypic variation. The ^{1}H NMR data, however, were generated using a 'global' method where indeed metabolites of a much larger number of classes may be detected that will share less biological pathways leading to phenotypic variation.

As already noted in the discussion of the results of our univariate analyses, interpretation of the results for the ^{1}H NMR data was often difficult due to the inherent properties of this metabolomics method. Nevertheless, we suspect that in future studies, hierarchical clustering on the basis of the genetic correlations among different peaks might be useful to reveal genetic relationships among different metabolites.

5.4.4 Quantitative genetic analyses of metabolomics data as reported in the literature

Other authors have given heritability estimates for metabolites as well. However, with the exception of a study by Shah and colleagues,[161] in all publications that we are aware of, the number of metabolites studied was too small and/or the resolution was too low (*e.g.*, all triglycerides lumped into one summary measure denoted "total triglyceride concentration") to denote the phenotypic data as "metabolomics data".[179–184] Furthermore, of note, a graphical example of variance components analysis of twin metabolomics data is given in Rahmioğlu *et al.* (their Figure 4).[185]

The reported study that is probably the closest to our current study is the one by Shah and co-workers.[161] In that study, quantitative measurements of 66 metabolites were performed belonging to acylcarnitine species, amino acids and free fatty acids, in blood plasma from eight (nontwin) families (total 117 individuals) heavily burdened with premature coronary artery disease. Univariate heritabilities were computed with methods that are equivalent to those employed in this chapter; however, the authors did not perform multivariate quantitative genetic analyses. Consistent with our findings, the authors report heritabilities within a large range over all investigated metabolites: for the metabolites for which the heritability estimate was statistically significant, they found heritabilities within the range [0.23–0.82].

In another interesting study, Pilia *et al.*[176] analyzed data on 98 quantitative traits relevant for cardiovascular function and personality, measured in a large cohort of Sardinians. These traits included levels of 'conventional metabolites' on the basis of clinical chemical measurements. In accordance with the current study, the authors used quantitative genetic methods in a "multistep multivariate" fashion to assess the genetic and environmental components of the phenotypic variances and covariances.

5.5 Conclusions

We have presented the results of a pilot investigation into the relative contribution of genetic variation to the variation observed in human blood plasma metabolite levels. Our analyses were based on the data obtained with two frequently used metabolomics platforms, *i.e.* LC–MS and ^{1}H NMR spectroscopy.

Notably, univariate quantitative genetic analysis of the lipid LC–MS data revealed a remarkable pattern in the heritabilities of TGs with different numbers of C-atoms and/or different numbers of double bonds in the fatty acids that may warrant further biochemical investigation. In multivariate analysis we found genetic covariance among lipids from the same lipid class (LC–MS). Therefore, we envision that the methods employed in this study can be used to discover novel biological pathways on the basis of "omics" type data obtained in families.

Due to the inherent properties of ^{1}H NMR, interpretation of the results based on these data was difficult. However, in general we found higher genetic covariance observed among variables observed with the 'targeted' lipid LC–MS platform with respect to those observed with the 'global' ^{1}H NMR metabolomics platform.

In conclusion, our study has demonstrated the use of uni- and multivariate quantitative genetic analysis to elucidate the importance of genetic variation to quantitative variation observed in human blood plasma metabolites. The statistical significance of our findings should be enhanced by replication in a larger cohort of families.

5.6 Acknowledgments

We thank all the twins and siblings who participated in this study. We gratefully acknowledge dr. MC Neale (Virginia Commonwealth University, VA, USA) and dr. MHM de Moor (VU University Amsterdam, Amsterdam, The Netherlands) for assistance with the OpenMx software. Furthermore we would like to acknowledge support from the Netherlands Bioinformatics Centre (NBIC) through its research programme BioRange (project number: SP 3.3.1); the Netherlands Metabolomics Centre; Spinozapremie NWO/SPI 56-464-14192; the Center for Medical Systems Biology (CMSB); Twin-family database for behavior genetics and genomics studies (NWO-MaGW 480-04-004) and NWO-MaGW Vervangingsstudie (NWO no. 400-05-717).

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.

Bibliography

1. Boomsma, D, Busjahn, A, and Peltonen, L. Classical twin studies and beyond. *Nat.Rev.Genet.* 2002:3(11), 872–882.
2. Dunn, W, Broadhurst, D, Atherton, H, Goodacre, R, and Griffin, J. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem.Soc.Rev.* 2011:40, 387–426.
3. Naso, PO. *Metamorphoses*. AD 8.
4. van der Greef, J and Smilde, A. Symbiosis of chemometrics and metabolomics: past, present, and future. *J.Chemometrics* 2005:19, 376–386.
5. Orešič, M. Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. *Nutr.Metab.Cardiovasc.Dis.* 2009:19(11), 816–824.
6. Crick, F. On protein synthesis. *Symp.Soc.Exp.Biol.* 1958:12, 138–163.
7. Crick, F. Central dogma of molecular biology. *Nature* 1970:227(5258), 561–563.
8. Schreiber, S. Small molecules: the missing link in the central dogma. *Nat.Chem.Biol.* 2005:1(2), 64–66.
9. Goodacre, R. Metabolomics of a superorganism. *J.Nutr.* 2007:137(1 Suppl), 259S–266S.
10. Gottesman, I and Gould, T. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am.J.Psychiatry* 2003:160(4), 636–645.
11. Comuzzie, A, Funahashi, T, Sonnenberg, G, Martin, L, Jacob, H, Black, A, Maas, D, Takahashi, M, Kihara, S, Tanaka, S, Matsuzawa, Y, Blangero, J, Cohen, D, and Kissebah, A. The genetic basis of plasma variation in adiponectin, a global endophenotype for obesity and the metabolic syndrome. *J.Clin.Endocrinol.Metab* 2001:86(9), 4321–4325.
12. Gieger, C, Geistlinger, L, Altmaier, E, de Hrabe, AM, Kronenberg, F, Meitinger, T, Mewes, H, Wichmann, H, Weinberger, K, Adamski, J, Illig,

T, and Suhre, K. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS.Genet.* 2008: 4(11), e1000282.
13. Henry, C. New 'ome' in town. *Chemical & Engineering Archives* 2002: 80(48), 66–70.
14. Lederberg, J and McCray, A. 'Ome sweet 'omics – a genealogical treasury of words. *The Scientist* 2001:15(7), 8.
15. van der Werf, M, Jellema, R, and Hankemeier, T. Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *J.Ind.Microbiol.Biotechnol.* 2005:32(6), 234–252.
16. Fiehn, O, Kopka, J, Dormann, P, Altmann, T, Trethewey, R, and Willmitzer, L. Metabolite profiling for plant functional genomics. *Nat.Biotechnol.* 2000:18(11), 1157–1161.
17. Nicholls, A, Mortishire-Smith, R, and Nicholson, J. NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. *Chem.Res.Toxicol.* 2003:16(11), 1395–1404.
18. Kell, D. Metabolomic biomarkers: search, discovery and validation. *Expert.Rev.Mol.Diagn.* 2007:7(4), 329–333.
19. Koek, M. *Gas chromatography mass spectrometry: key technology in metabolomics*. Ph.D. thesis, Leiden University, Leiden, The Netherlands, 2009.
20. Brown, M, Dunn, W, Ellis, D, Goodacre, R, Handl, J, Knowles, J, O'Hagan, S, Spasic, I, and Kell, D. A metabolome pipeline: from concept to data to knowledge. *Metabolomics* 2005:1(1), 39–51.
21. Scalbert, A, Brennan, L, Fiehn, O, Hankemeier, T, Kristal, B, van Ommen, B, Pujos-Guillot, E, Verheij, E, Wishart, D, and Wopereis, S. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics* 2009:5(4), 435–458.
22. Craig, A, Cloarec, O, Holmes, E, Nicholson, J, and Lindon, J. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal.Chem.* 2006:78(7), 2262–2267.
23. Jolliffe, I. *Principal component analysis*. Springer-Verlag New York, Inc., New York, NY, USA, 2nd edition, 2002.
24. Westerhuis, J, Hoefsloot, H, Smit, S, Vis, D, Smilde, A, van Velzen, E, van Duijnhoven, J, and van Dorsten, F. Assessment of PLSDA cross validation. *Metabolomics* 2008:4, 81–89.
25. Galton, F. *Hereditary genius: an inquiry into its laws and consequences*. MacMillan & Co., London / New York, 2nd edition, 1892.
26. Galton, F. *English men of science: their nature and nurture*. MacMillan & Co., London, 1874.
27. Neale, M and Cardon, L. *Methodology for genetic studies of twins and families*, volume 67. Kluwer Academic Publishers, Dordrecht, 1992.
28. Falconer, D. *Introduction to quantitative genetics*. Oliver & Boyd, Edinburgh / London, 2nd edition, 1961.

29. Medland, S and Hatemi, P. Political science, biometrical theory, and twin studies: a methodological introduction. *Political Analysis* 2009:17, 191–214.
30. Rijsdijk, F and Sham, P. Analytic approaches to twin data using structural equation models. *Brief.Bioinform.* 2002:3(2), 119–133.
31. Sung, J, Cho, S, Song, Y, Lee, K, Choi, E, Ha, M, Kim, J, Kim, H, Kim, Y, Shin, E, Kim, Y, Yoo, K, Park, C, and Kimm, K. Do we need more twin studies? The Healthy Twin Study, Korea. *Int.J.Epidemiol.* 2006: 35(2), 488–490.
32. Bollen, K. *Structural equations with latent variables.* John Wiley & Sons, New York, 1989.
33. Hox, J and Bechger, T. An introduction to structural equation modeling. *Family Science Review* 1998:11, 354–373.
34. Wright, S. Correlation and causation. *Journal of Agricultural Research* 1921:20, 557–585.
35. Jinks, J and Fulker, D. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychol.Bull.* 1970:73(5), 311–349.
36. Lay, D. *Linear algebra and its applications.* Addison-Wesley, Boston, 3rd edition, 2003.
37. Visscher, P, Hill, W, and Wray, N. Heritability in the genomics era – concepts and misconceptions. *Nat.Rev.Genet.* 2008:9(4), 255–266.
38. Posthuma, D, Beem, A, de Geus, E, van Baal, G, von Hjelmborg, J, Iachine, I, and Boomsma, D. Theory and practice in quantitative genetics. *Twin Research* 2003:6(5), 361–376.
39. Evans, D, Gillespie, N, and Martin, N. Biometrical genetics. *Biol.Psychol.* 2002:61(1-2), 33–51.
40. Dolan, C. *Biometric decomposition of phenotypic means in human samples.* Ph.D. thesis, University of Amsterdam, Amsterdam, The Netherlands, 1992.
41. Kaufman, L and Rousseeuw, P. *Finding groups in data – an introduction to cluster analysis.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.
42. Kriegel, H, Kröger, P, and Zimek, A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on knowledge discovery from data* 2009: 3(1, article 1 (58 pages)).
43. Friedman, J and Meulman, J. Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society Series B* 2004:66, 815–849.
44. Davidov, E, Holland, J, Marple, E, and Naylor, S. Advancing drug discovery through systems biology. *Drug.Discov.Today* 2003:8(4), 175–183.
45. Clish, C, Davidov, E, Orešič, M, Plasterer, T, Lavine, G, Londo, T, Meys, M, Snell, P, Stochaj, W, Adourian, A, Zhang, X, Morel, N, Neumann, E, Verheij, E, Vogels, J, Havekes, L, Afeyan, N, Regnier, F, van der Greef, J, and Naylor, S. Integrative biological analysis of the APOE*3-leiden transgenic mouse. *OMICS* 2004:8(1), 3–13.

46. Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 1918:52, 399–433.
47. Plomin, R, Haworth, C, and Davis, O. Common disorders are quantitative traits. *Nat.Rev.Genet.* 2009:10(12), 872–878.
48. Frazer, K, Murray, S, Schork, N, and Topol, E. Human genetic variation and its contribution to complex traits. *Nat.Rev.Genet.* 2009:10(4), 241–251.
49. McCarthy, M, Abecasis, G, Cardon, L, Goldstein, D, Little, J, Ioannidis, J, and Hirschhorn, J. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat.Rev.Genet.* 2008:9(5), 356–369.
50. Pearson, T and Manolio, T. How to interpret a genome-wide association study. *JAMA* 2008:299(11), 1335–1344.
51. Illig, T, Gieger, C, Zhai, G, Romisch-Margl, W, Wang-Sattler, R, Prehn, C, Altmaier, E, Kastenmuller, G, Kato, B, Mewes, H, Meitinger, T, de Angelis, M, Kronenberg, F, Soranzo, N, Wichmann, H, Spector, T, Adamski, J, and Suhre, K. A genome-wide perspective of genetic variation in human metabolism. *Nat.Genet.* 2010:42(2), 137–141.
52. Hicks, A, Pramstaller, P, Johansson, Å, Vitart, V, Rudan, I, Ugocsai, P, Aulchenko, Y, Franklin, C, Liebisch, G, Erdmann, J, Jonasson, I, Zorkoltseva, I, Pattaro, C, Hayward, C, Isaacs, A, Hengstenberg, C, Campbell, S, Gnewuch, C, Janssens, A, Kirichenko, A, König, I, Marroni, F, Polasek, O, Demirkan, A, Kolcic, I, Schwienbacher, C, Igl, W, Biloglav, Z, Witteman, J, Pichler, I, Zaboli, G, Axenovich, T, Peters, A, Schreiber, S, Wichmann, H, Schunkert, H, Hastie, N, Oostra, B, Wild, S, Meitinger, T, Gyllensten, U, van Duijn, C, Wilson, J, Wright, A, Schmitz, G, and Campbell, H. Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS.Genet.* 2009:5(10), e1000672.
53. Lee, S, van der Werf, J, Hayes, B, Goddard, M, and Visscher, P. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS.Genet.* 2008:4(10), e1000231.
54. Tanaka, T, Shen, J, Abecasis, G, Kisialiou, A, Ordovas, J, Guralnik, J, Singleton, A, Bandinelli, S, Cherubini, A, Arnett, D, Tsai, M, and Ferrucci, L. Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS.Genet.* 2009:5(1), e1000338.
55. Goring, H, Curran, J, Johnson, M, Dyer, T, Charlesworth, J, Cole, S, Jowett, J, Abraham, L, Rainwater, D, Comuzzie, A, Mahaney, M, Almasy, L, MacCluer, J, Kissebah, A, Collier, G, Moses, E, and Blangero, J. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat.Genet.* 2007:39(10), 1208–1216.
56. Emilsson, V, Thorleifsson, G, Zhang, B, Leonardson, A, Zink, F, Zhu, J, Carlson, S, Helgason, A, Walters, G, Gunnarsdottir, S, Mouy, M, Steinthorsdottir, V, Eiriksdottir, G, Bjornsdottir, G, Reynisdottir, I, Gudbjartsson, D, Helgadottir, A, Jonasdottir, A, Jonasdottir, A, Styrkarsdot-

tir, U, Gretarsdottir, S, Magnusson, K, Stefansson, H, Fossdal, R, Kristjansson, K, Gislason, H, Stefansson, T, Leifsson, B, Thorsteinsdottir, U, Lamb, J, Gulcher, J, Reitman, M, Kong, A, Schadt, E, and Stefansson, K. Genetics of gene expression and its effect on disease. *Nature* 2008: 452(7186), 423–428.

57. Manolio, T, Collins, F, Cox, N, Goldstein, D, Hindorff, L, Hunter, D, McCarthy, M, Ramos, E, Cardon, L, Chakravarti, A, Cho, J, Guttmacher, A, Kong, A, Kruglyak, L, Mardis, E, Rotimi, C, Slatkin, M, Valle, D, Whittemore, A, Boehnke, M, Clark, A, Eichler, E, Gibson, G, Haines, J, Mackay, T, McCarroll, S, and Visscher, P. Finding the missing heritability of complex diseases. *Nature* 2009:461(7265), 747–753.
58. Maher, B. Personal genomes: The case of the missing heritability. *Nature* 2008:456(7218), 18–21.
59. Manolio, T. Genomewide association studies and assessment of the risk of disease. *N.Engl.J.Med.* 2010:363(2), 166–176.
60. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* 2008:456(7223), 728–731.
61. Visscher, PM, Yang, J, and Goddard, ME. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res.Hum.Genet.* 2010:13(6), 517–524.
62. Yang, J, Benyamin, B, McEvoy, BP, Gordon, S, Henders, AK, Nyholt, DR, Madden, PA, Heath, AC, Martin, NG, Montgomery, GW, Goddard, ME, and Visscher, PM. Common SNPs explain a large proportion of the heritability for human height. *Nat.Genet.* 2010:42(7), 565–569.
63. Manolio, T, Brooks, L, and Collins, F. A HapMap harvest of insights into the genetics of common disease. *J.Clin.Invest.* 2008:118(5), 1590–1605.
64. Sebastiani, P, Timofeev, N, Dworkis, D, Perls, T, and Steinberg, M. Genome-wide association studies and the genetic dissection of complex traits. *Am.J.Hematol.* 2009:84(8), 504–515.
65. Clarke, A and Cooper, D. GWAS: heritability missing in action? *Eur.J.Hum.Genet.* 2010:18(8), 859–861.
66. Hardy, J and Singleton, A. Genomewide association studies and human disease. *N.Engl.J.Med.* 2009:360(17), 1759–1768.
67. Bourgain, C, Genin, E, Cox, N, and Clerget-Darpoux, F. Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases? *Eur.J.Hum.Genet.* 2007:15(3), 260–263.
68. Thomas, D. Gene-environment-wide association studies: emerging approaches. *Nat.Rev.Genet.* 2010:11(4), 259–272.
69. Martin, N, Boomsma, D, and Machin, G. A twin-pronged attack on complex traits. *Nat.Genet.* 1997:17(4), 387–392.
70. Fischer, K, Bot, A, Zwaan, B, and Brakefield, P. Genetic and environmental sources of egg size variation in the butterfly *Bicyclus anynana*. *Heredity* 2004:92(3), 163–169.
71. van der Greef, J, Davidov, E, Verheij, E, Vogels, J, van der Heijden, R, Adourian, A, Oresic, M, Marple, E, and Naylor, S. The role of metabolom-

ics in systems biology. In Harrigan, G and Goodacre, R, editors, *Metabolic profiling: its role in biomarker discovery and gene function analysis*, chapter 11, pages 171–198. Kluwer Academic Publishers, Boston: 2004.

72. Nanki, T, Kohsaka, H, Mizushima, N, Ollier, W, Carson, D, and Miyasaka, N. Genetic control of T cell receptor BJ gene expression in peripheral lymphocytes of normal and rheumatoid arthritis monozygotic twins. *J.Clin.Invest.* 1996:98(7), 1594–1601.
73. German, J, Roberts, M, and Watkins, S. Personal metabolomics as a next generation nutritional assessment. *J.Nutr.* 2003:133(12), 4260–4266.
74. Koek, M, Muilwijk, B, van der Werf, M, and Hankemeier, T. Microbial metabolomics with gas chromatography/mass spectrometry. *Anal.Chem.* 2006:78(4), 1272–1281.
75. Murphy, D. The biogenesis and functions of lipid bodies in animals, plants and microorganisms. *Prog.Lipid.Res.* 2001:40(5), 325–438.
76. Baur, L, O'Connor, J, Pan, D, Wu, B, O'Connor, M, and Storlien, L. Relationships between the fatty acid composition of muscle and erythrocyte membrane phospholipid in young children and the effect of type of infant feeding. *Lipids* 2000:35(1), 77–82.
77. Swift, RW. The effects of low environmental temperature upon metabolism: II. The influence of shivering, subcutaneous fat, and skin temperature on heat production. *J.Nutr.* 1932:5, 227–249.
78. Dodds, P. Incorporation of xenobiotic carboxylic acids into lipids. *Life Sci.* 1991:49(9), 629–649.
79. Tew, D, Southan, C, Rice, S, Lawrence, M, Li, H, Boyd, H, Moores, K, Gloger, I, and Macphee, C. Purification, properties, sequencing, and cloning of a lipoprotein-associated, serine-dependent phospholipase involved in the oxidative modification of low-density lipoproteins. *Arterioscler.Thromb.Vasc.Biol.* 1996:16(4), 591–599.
80. Elstad, M, Stafforini, D, McIntyre, T, Prescott, S, and Zimmerman, G. Platelet-activating factor acetylhydrolase increases during macrophage differentiation. A novel mechanism that regulates accumulation of platelet-activating factor. *J.Biol.Chem.* 1989:264(15), 8467–8470.
81. Quinn, M, Parthasarathy, S, and Steinberg, D. Lysophosphatidylcholine: a chemotactic factor for human monocytes and its potential role in atherogenesis. *Proc.Natl.Acad.Sci.USA* 1988:85(8), 2805–2809.
82. Glomset, J. The plasma lecithins:cholesterol acyltransferase reaction. *J.Lipid Res.* 1968:9(2), 155–167.
83. Kern, H, Volk, T, Knauer-Schiefer, S, Mieth, T, Rustow, B, Kox, W, and Schlame, M. Stimulation of monocytes and platelets by short-chain phosphatidylcholines with and without terminal carboxyl group. *Biochim.Biophys.Acta* 1998:1394(1), 33–42.
84. Coleman, R. Biochemistry of bile secretion. *Biochem.J.* 1987:244(2), 249–261.
85. Zachowski, A. Phospholipids in animal eukaryotic membranes: transverse asymmetry and movement. *Biochem.J.* 1993:294 (Pt 1), 1–14.

86. McKeone, B, Patsch, J, and Pownall, H. Plasma triglycerides determine low density lipoprotein composition, physical properties, and cell-specific binding in cultured cells. *J.Clin.Invest.* 1993:91(5), 1926–1933.
87. Sigal, Y, McDermott, M, and Morris, A. Integral membrane lipid phosphatases/phosphotransferases: common structure and diverse functions. *Biochem.J.* 2005:387(Pt 2), 281–293.
88. Veldhuizen, R, Nag, K, Orgeig, S, and Possmayer, F. The role of lipids in pulmonary surfactant. *Biochim.Biophys.Acta* 1998:1408(2-3), 90–108.
89. Acton, S, Rigotti, A, Landschulz, K, Xu, S, Hobbs, H, and Krieger, M. Identification of scavenger receptor SR-BI as a high density lipoprotein receptor. *Science* 1996:271(5248), 518–520.
90. Boomsma, D, de Geus, E, Vink, J, Stubbe, J, Distel, M, Hottenga, J, Posthuma, D, van Beijsterveldt, T, Hudziak, J, Bartels, M, and Willemsen, G. Netherlands Twin Register: from twins to twin families. *Twin.Res.Hum.Genet.* 2006:9(6), 849–857.
91. Hoekstra, R, Bartels, M, and Boomsma, D. Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. *Learning and individual differences* 2007:17, 97–114.
92. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.
93. Vandeginste, B, Massart, D, Buydens, L, de Jong, S, Lewi, P, and Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics: part B*. Elsevier, Amsterdam, 1998.
94. Barnes, R, Dhanoa, M, and Lister, S. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy* 1989:43, 772–777.
95. Sokal, R. Distance as a measure of taxonomic similarity. *Systematic zoology* 1961:10, 70–79.
96. Young, G and Householder, A. Discussion of a set of points in terms of their mutual distances. *Psychometrika* 1938:3, 19–22.
97. Kruskal, WH and Wallis, W. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952:47(260), 583–621.
98. Hochberg, Y and Tamhane, A. *Multiple comparison procedures*. John Wiley & Sons, New York, 1987.
99. Sokal, R and Rohlf, F. The comparison of dendrograms by objective methods. *Taxon* 1962:11(2), 33–40.
100. Sneath, P and Sokal, R. *Numerical taxonomy: the principles and practice of numerical classification*. W.H. Freeman & Co., New York, 1973.
101. Suzuki, R and Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006:22(12), 1540–1542.
102. Damian, D, Orešič, M, Verheij, E, Meulman, J, Friedman, J, Adourian, A, Morel, N, Smilde, A, and van der Greef, J. Applications of a new subspace

clustering algorithm (COSA) in medical systems biology. *Metabolomics* 2007:3, 629–649.

103. Tan, Q, Christensen, K, Christiansen, L, Frederiksen, H, Bathum, L, Dahlgaard, J, and Kruse, T. Genetic dissection of gene expression observed in whole blood samples of elderly Danish twins. *Hum.Genet.* 2005: 117(2-3), 267–274.
104. Omori-Inoue, M, Fukata, H, Komiyama, M, Todaka, E, Osada, H, Aburatani, H, and Mori, C. The contamination levels of organochlorines and the pattern of gene expressions in human umbilical cords from intra-pairs of twins at delivery. *Reprod.Toxicol.* 2007:23(3), 283–289.
105. Zhou, X, Tan, F, Xiong, M, Arnett, F, and Feghali-Bostwick, C. Monozygotic twins clinically discordant for scleroderma show concordance for fibroblast gene expression profiles. *Arthritis Rheum.* 2005:52(10), 3305–3314.
106. Matigian, N, Windus, L, Smith, H, Filippich, C, Pantelis, C, McGrath, J, Mowry, B, and Hayward, N. Expression profiling in monozygotic twins discordant for bipolar disorder reveals dysregulation of the WNT signalling pathway. *Mol.Psychiatry* 2007:12(9), 815–825.
107. Teuffel, O, Betts, D, Dettling, M, Schaub, R, Schafer, B, and Niggli, F. Prenatal origin of separate evolution of leukemia in identical twins. *Leukemia* 2004:18(10), 1624–1629.
108. Tsang, T, Huang, J, Holmes, E, and Bahn, S. Metabolic profiling of plasma from discordant schizophrenia twins: correlation between lipid signals and global functioning in female schizophrenia patients. *J.Proteome Res.* 2006:5(4), 756–760.
109. Pietiläinen, K, Sysi-Aho, M, Rissanen, A, Seppanen-Laakso, T, Yki-Jarvinen, H, Kaprio, J, and Orešič, M. Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects – a monozygotic twin study. *PLoS ONE* 2007:2, e218.
110. Hastie, T, Tibshirani, R, and Friedman, J. *The elements of statistical learning; data mining, inference, and prediction.* Springer, New York, 2001.
111. Iselius, L. Analysis of family resemblance for lipids and lipoproteins. *Clin.Genet.* 1979:15(4), 300–306.
112. Snieder, H, van Doornen, L, and Boomsma, D. Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies. *Arterioscler.Thromb.Vasc.Biol.* 1999:19(12), 2826–2834.
113. Boomsma, D, Kempen, H, Gevers Leuven, J, Havekes, L, de Knijff, P, and Frants, R. Genetic analysis of sex and generation differences in plasma lipid, lipoprotein, and apolipoprotein levels in adolescent twins and their parents. *Genet.Epidemiol.* 1996:13(1), 49–60.
114. Cohn, J, McNamara, J, Cohn, S, Ordovas, J, and Schaefer, E. Postprandial plasma lipoprotein changes in human subjects of different ages. *J.Lipid Res.* 1988:29(4), 469–479.
115. Steinmetz, V, Sévilla, F, and Bellon-Maurel, V. A methodology for sensor

fusion design: application to fruit quality assessment. *J.Agric.Engng.Res.* 1999:74, 21–31.

116. Smilde, A, van der Werf, M, Bijlsma, S, van der Werff-van der Vat BJ, and Jellema, R. Fusion of mass spectrometry-based metabolomics data. *Anal.Chem.* 2005:77(20), 6729–6736.
117. Zelena, E, Dunn, W, Broadhurst, D, Francis-McIntyre, S, Carroll, K, Begley, P, O'Hagan, S, Knowles, J, Halsall, A, Wilson, I, and Kell, D. Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Anal.Chem.* 2009:81(4), 1357–1364.
118. Feudale, R, Woody, N, Tan, H, Myles, A, Brown, S, and Ferré, J. Transfer of multivariate calibration models: a review. *Chemom.Intell.Lab.Syst.* 2002:64, 181–192.
119. Alam, T, Alam, M, McIntyre, S, Volk, D, Neerathilingam, M, and Luxon, B. Investigation of chemometric instrumental transfer methods for high-resolution NMR. *Anal.Chem.* 2009:81(11), 4433–4443.
120. Keun, H, Ebbels, T, Antti, H, Bollard, M, Beckonert, O, Schlotterbeck, G, Senn, H, Niederhauser, U, Holmes, E, Lindon, J, and Nicholson, J. Analytical reproducibility in 1H NMR-based metabonomic urinalysis. *Chem.Res.Toxicol.* 2002:15(11), 1380–1386.
121. Dumas, M, Maibaum, E, Teague, C, Ueshima, H, Zhou, B, Lindon, J, Nicholson, J, Stamler, J, Elliott, P, Chan, Q, and Holmes, E. Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. *Anal.Chem.* 2006:78(7), 2199–2208.
122. Sangster, T, Major, H, Plumb, R, Wilson, A, and Wilson, I. A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. *Analyst* 2006:131(10), 1075–1078.
123. Gika, H, Theodoridis, G, Wingate, J, and Wilson, I. Within-day reproducibility of an HPLC-MS-based method for metabonomic analysis: application to human urine. *J.Proteome.Res.* 2007:6(8), 3291–3303.
124. Theodoridis, G, Gika, H, and Wilson, I. LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics. *Trends in Analytical Chemistry* 2008:27(3), 251–260.
125. Burton, L, Ivosev, G, Tate, S, Impey, G, Wingate, J, and Bonner, R. Instrumental and experimental effects in LC-MS based metabolomics. *J.Chrom.B* 2008:871, 227–235.
126. Dunn, W, Broadhurst, D, Brown, M, Baker, P, Redman, C, Kenny, L, and Kell, D. Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system. *J.Chrom.B* 2008:871, 288–298.
127. Bijlsma, S, Bobeldijk, I, Verheij, E, Ramaker, R, Kochhar, S, Macdonald, I, van Ommen, B, and Smilde, A. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal.Chem.* 2006:78(2), 567–574.

128. van der Kloet, F, Jellema, R, Verheij, E, and Bobeldijk, I. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *J Proteome Res.* 2009:8(11), 5132–5141.
129. Kolen, M and Jarjoura, D. Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrika* 1987:52(1), 43–59.
130. Van der Linden, W. A test-theoretic approach to observed-score equating. *Psychometrika* 2000:65(4), 437–456.
131. Wagner, S, Scholz, K, Sieber, M, Kellert, M, and Voelkel, W. Tools in metabonomics: an integrated validation approach for LC-MS metabolic profiling of mercapturic acids in human urine. *Anal.Chem.* 2007:79(7), 2918–2926.
132. Vogels, J, Tas, A, Venekamp, J, and van der Greef, J. Partial linear fit: a new NMR spectroscopy preprocessing tool for pattern recognition applications. *J.Chemometrics* 1996:10, 425–438.
133. Hendriks, M, Smit, S, Akkermans, W, Reijmers, T, Eilers, P, Hoefsloot, H, Rubingh, C, de Koster, C, Aerts, J, and Smilde, A. How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. *Proteomics.* 2007:7(20), 3672–3680.
134. Bolstad, B. Probe level quantile normalization of high density oligonucleotide array data. http://bmbolstad.com/stuff/qnorm.pdf, 2001.
135. Bolstad, B, Irizarry, R, Astrand, M, and Speed, T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003:19(2), 185–193.
136. Cleveland, W. *The elements of graphing data.* Hobart Press, Summit, N.J., 2nd edition, 1994.
137. Wilk, M and Gnanadesikan, R. Probability plotting methods for analysis of data. *Biometrika* 1968:55(1), 1–17.
138. Cawley, S, Bekiranov, S, Ng, H, Kapranov, P, Sekinger, E, Kampa, D, Piccolboni, A, Sementchenko, V, Cheng, J, Williams, A, Wheeler, R, Wong, B, Drenkow, J, Yamanaka, M, Patel, S, Brubaker, S, Tammana, H, Helt, G, Struhl, K, and Gingeras, T. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 2004:116(4), 499–509.
139. Kim, J, Tchernyshyov, I, Semenza, G, and Dang, C. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metab.* 2006:3(3), 177–185.
140. Chen, H, Yu, S, Chen, C, Chang, G, Chen, C, Yuan, A, Cheng, C, Wang, C, Terng, H, Kao, S, Chan, W, Li, H, Liu, C, Singh, S, Chen, W, Chen, J, and Yang, P. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N.Engl.J.Med.* 2007:356(1), 11–20.
141. Higgs, R, Knierman, M, Gelfanova, V, Butler, J, and Hale, J. Comprehensive label-free method for the relative quantification of proteins from biological samples. *J.Proteome.Res* 2005:4(4), 1442–1450.
142. Angoff, W. Scales, norms, and equivalent scores. In Thorndike, R, editor,

Educational measurement, pages 562–600. American Council on Education, Washington, D.C., 2nd edition: 1971.

143. Gentleman, R, Carey, V, Bates, D, Bolstad, B, Dettling, M, Dudoit, S, Ellis, B, Gautier, L, Ge, Y, Gentry, J, Hornik, K, Hothorn, T, Huber, W, Iacus, S, Irizarry, R, Leisch, F, Li, C, Maechler, M, Rossini, A, Sawitzki, G, Smith, C, Smyth, G, Tierney, L, Yang, J, and Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004:5(10), R80.
144. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
145. Jouan-Rimbaud, D, Massart, D, Saby, C, and Puel, C. Determination of the representativity between two multidimensional data sets by a comparison of their structure. *Chemom.Intell.Lab.Syst.* 1998:40, 129–144.
146. Frisby, J and Clatworthy, J. Learning to see complex random-dot stereograms. *Perception* 1975:4, 173–178.
147. Deprez, S, Sweatman, B, Connor, S, Haselden, J, and Waterfield, C. Optimisation of collection, storage and preparation of rat plasma for 1H NMR spectroscopic analysis in toxicology studies to determine inherent variation in biochemical profiles. *J.Pharm.Biomed.Anal.* 2002:30(4), 1297–1310.
148. Mehr, K, John, B, Russell, D, and Avizonis, D. Electronic referencing techniques for quantitative NMR: pitfalls and how to avoid them using amplitude-corrected referencing through signal injection. *Anal.Chem.* 2008:80(21), 8320–8323.
149. Lauridsen, M, Hansen, S, Jaroszewski, J, and Cornett, C. Human urine as test material in 1H NMR-based metabonomics: recommendations for sample preparation and storage. *Anal.Chem.* 2007:79(3), 1181–1186.
150. Srinivasan, R and Stewart, R. The catalysis of proton exchange in creatinine by general acids and general bases. *Can.J.Chem.* 1975:53, 224–231.
151. Mackay, T. The genetic architecture of quantitative traits. *Annu.Rev.Genet.* 2001:35, 303–339.
152. Eaves, L. Putting the 'human' back in genetics: modeling the extended kinships of twins. *Twin Res.Hum.Genet.* 2009:12(1), 1–7.
153. Hu, C, van der Heijden, R, Wang, M, van der Greef, J, Hankemeier, T, and Xu, G. Analytical strategies in lipidomics and applications in disease biomarker discovery. *Journal of Chromatography B* 2009:877, 2836–2846.
154. Vrije Universiteit - Nederlands Tweelingen Register. http://www.tweelingenregister.org/. Accessed 27-March-2011.
155. Draisma, H, Reijmers, T, Bobeldijk-Pastorova, I, Meulman, J, Estourgie-Van Burk, G, Bartels, M, Ramaker, R, van der Greef, J, Boomsma, D, and Hankemeier, T. Similarities and differences in lipidomics profiles among healthy monozygotic twin pairs. *OMICS* 2008:12(1), 17–31.
156. Draisma, H, Reijmers, T, van der Kloet, F, Bobeldijk-Pastorova, I, Spies-Faber, E, Vogels, J, Meulman, J, Boomsma, D, Van der Greef, J, and Han-

kemeier, T. Equating, or correction for between-block effects with application to body fluid LC-MS and NMR metabolomics data sets. *Anal.Chem.* 2010:82(3), 1039–1046.

157. Willemsen, G, de Geus, E, Bartels, M, van Beijsterveldt, C, Brooks, A, Estourgie-Van Burk, G, Fugman, D, Hoekstra, C, Hottenga, J, Kluft, K, Meijer, P, Montgomery, G, Rizzu, P, Sondervan, D, Smit, A, Spijker, S, Suchiman, H, Tischfield, J, Lehner, T, Slagboom, P, and Boomsma, D. The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res.Hum.Genet.* 2010:13(3), 231–245.
158. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.
159. Clyne, B and Olshaker, J. The C-reactive protein. *J.Emerg.Med.* 1999: 17(6), 1019–1025.
160. Scriver, C. Garrod's foresight; our hindsight. *J.Inherit.Metab Dis.* 2001: 24(2), 93–116.
161. Shah, S, Hauser, E, Bain, J, Muehlbauer, M, Haynes, C, Stevens, R, Wenner, B, Dowdy, Z, Granger, C, Ginsburg, G, Newgard, C, and Kraus, W. High heritability of metabolomic profiles in families burdened with premature cardiovascular disease. *Mol.Syst.Biol.* 2009:5, 258.
162. Vaidyanathan, S, Harrigan, G, and Goodacre, R. Introduction. In Vaidyanathan, S, Harrigan, G, and Goodacre, R, editors, *Metabolome analyses: strategies for systems biology*, chapter 1, pages 1–8. Springer Science+Business Media, Inc., New York, NY: 2005.
163. van der Greef, J, Martin, S, Juhasz, P, Adourian, A, Plasterer, T, Verheij, E, and McBurney, R. The art and practice of systems biology in medicine: mapping patterns of relationships. *J.Proteome.Res.* 2007:6(4), 1540–1559.
164. Carey, G. Inference about genetic correlations. *Behav.Genet.* 1988:18(3), 329–338.
165. Schmitz, S, Cherny, S, and Fulker, D. Increase in power through multivariate analyses. *Behav.Genet.* 1998:28(5), 357–363.
166. Posthuma, D and Boomsma, D. A note on the statistical power in extended twin designs. *Behav.Genet.* 2000:30(2), 147–158.
167. Schmitt, J, Lenroot, R, Wallace, G, Ordaz, S, Taylor, K, Kabani, N, Greenstein, D, Lerch, J, Kendler, K, Neale, M, and Giedd, J. Identification of genetically mediated cortical networks: a multivariate study of pediatric twins and siblings. *Cerebral Cortex* 2008:18, 1737–1747.
168. Schmitt, J, Lenroot, R, Ordaz, S, Wallace, G, Lerch, J, Evans, A, Prom, E, Kendler, K, Neale, M, and Giedd, J. Variance decomposition of MRI-based covariance maps using genetically informative samples and structural equation modeling. *Neuroimage.* 2009:47(1), 56–64.
169. Boker, S, Neale, M, Maes, H, Wilde, M, Spiegel, M, Brick, T, Spies, J, Estabrook, R, Kenny, S, Bates, T, Mehta, P, and Fox, J. OpenMx: an open source extended structural equation modeling framework. *Psychometrika* : advance online publication 6 January 2011; doi:

10.1007/S11336-010-9200-6.
170. Nadder, T, Silberg, J, Eaves, L, Maes, H, and Meyer, J. Genetic effects on ADHD symptomatology in 7- to 13-year-old twins: results from a telephone survey. *Behav.Genet.* 1998:28(2), 83–99.
171. Neale, M and Miller, M. The use of likelihood-based confidence intervals in genetic models. *Behav.Genet.* 1997:27(2), 113–120.
172. Giedd, J, Schmitt, J, and Neale, M. Structural brain magnetic resonance imaging of pediatric twins. *Hum.Brain Mapp.* 2007:28(6), 474–481.
173. Baaré, W, Hulshoff Pol, H, Boomsma, D, Posthuma, D, de Geus, E, Schnack, H, van Haren, N, van Oel, C, and Kahn, R. Quantitative genetic modeling of variation in human brain morphology. *Cerebral Cortex* 2001: 11, 816–824.
174. Atchley, W, Plummer, A, and Riska, B. Genetics of mandible form in the mouse. *Genetics* 1985:111(3), 555–577.
175. Eyler, L, Prom-Wormley, E, Fennema-Notestine, C, Panizzon, M, Neale, M, Jernigan, T, Fischl, B, Franz, C, Lyons, M, Stevens, A, Pacheco, J, Perry, M, Schmitt, J, Spitzer, N, Seidman, L, Thermenos, H, Tsuang, M, Dale, A, and Kremen, W. Genetic patterns of correlation among subcortical volumes in humans: Results from a magnetic resonance imaging twin study. *Hum.Brain Mapp.* : advance online publication 22 June 2010; doi: 10.1002/hbm.21054.
176. Pilia, G, Chen, W, Scuteri, A, Orru, M, Albai, G, Dei, M, Lai, S, Usala, G, Lai, M, Loi, P, Mameli, C, Vacca, L, Deiana, M, Olla, N, Masala, M, Cao, A, Najjar, S, Terracciano, A, Nedorezov, T, Sharov, A, Zonderman, A, Abecasis, G, Costa, P, Lakatta, E, and Schlessinger, D. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS.Genet.* 2006:2(8), e132.
177. Voet, D, Voet, J, and Pratt, C. Synthesis and degradation of lipids. In *Principles of biochemistry*, chapter 20, pages 677–731. John Wiley & Sons, Hoboken,NJ, 3rd edition: 2008.
178. Nicholson, J and Lindon, J. Systems biology: Metabonomics. *Nature* 2008:455(7216), 1054–1056.
179. Fenger, M, Benyamin, B, Schousboe, K, Sørensen, T, and Kyvik, K. Variance decomposition of apolipoproteins and lipids in Danish twins. *Atherosclerosis* 2007:191(1), 40–47.
180. Beekman, M, Heijmans, B, Martin, N, Pedersen, N, Whitfield, J, DeFaire, U, van Baal, G, Snieder, H, Vogler, G, Slagboom, P, and Boomsma, D. Heritabilities of apolipoprotein and lipid levels in three countries. *Twin Res.* 2002:5(2), 87–97.
181. Kullo, I, de Andrade, M, Boerwinkle, E, McConnell, J, Kardia, S, and Turner, S. Pleiotropic genetic effects contribute to the correlation between HDL cholesterol, triglycerides, and LDL particle size in hypertensive sibships. *Am.J.Hypertens.* 2005:18(1), 99–103.
182. Mathias, R, Deepa, M, Deepa, R, Wilson, A, and Mohan, V. Heritability of quantitative traits associated with type 2 diabetes mellitus in large

multiplex families from South India. *Metabolism* 2009:58(10), 1439–1445.
183. Benyamin, B, Sørensen, T, Schousboe, K, Fenger, M, Visscher, P, and Kyvik, K. Are there common genetic and environmental factors behind the endophenotypes associated with the metabolic syndrome? *Diabetologica* 2007:50, 1880–1888.
184. Rahman, I, Bennet, A, Pedersen, N, de Faire, U, Svensson, P, and Magnusson, P. Genetic dominance influences blood biomarker levels in a sample of 12,000 Swedish elderly twins. *Twin Res.Hum.Genet.* 2009:12(3), 286–294.
185. Rahmioğlu, N and Ahmadi, K. Classical twin design in modern pharmacogenomics studies. *Pharmacogenomics* 2010:11(2), 215–226.
186. Demetrashvili, M, Kron, K, Pethe, V, Bapat, B, and Briollais, L. How to deal with batch effect in sequential microarray experiments? *Molecular Informatics* 2010:29, 387–393.
187. Sims, A, Smethurst, G, Hey, Y, Okoniewski, M, Pepper, S, Howell, A, Miller, C, and Clarke, R. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Med.Genomics* 2008:1, 42.
188. Benito, M, Parker, J, Du, Q, Wu, J, Xiang, D, Perou, C, and Marron, J. Adjustment of systematic microarray data biases. *Bioinformatics* 2004: 20(1), 105–114.
189. Jiang, H, Deng, Y, Chen, H, Tao, L, Sha, Q, Chen, J, Tsai, C, and Zhang, S. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 2004:5, 81.
190. Kim, K, Ki, D, Jeong, H, Jeung, H, Chung, H, and Rha, S. Novel and simple transformation algorithm for combining microarray data sets. *BMC Bioinformatics* 2007:8, 218.
191. Johnson, W, Li, C, and Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007: 8(1), 118–127.
192. Scherer, A, editor. *Batch effects and noise in microarray experiments.* John Wiley & Sons, Ltd., Chichester, West Sussex, UK, 1st edition, 2009.
193. Shabalin, A, Tjelmeland, H, Fan, C, Perou, C, and Nobel, A. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008:24(9), 1154–1160.
194. Begley, P, Francis-McIntyre, S, Dunn, W, Broadhurst, D, Halsall, A, Tseng, A, Knowles, J, Goodacre, R, and Kell, D. Development and performance of a gas chromatography-time-of-flight mass spectrometry analysis for large-scale nontargeted metabolomic studies of human serum. *Anal.Chem.* 2009:81(16), 7038–7046.
195. Fiehn, O, Kristal, B, van Ommen, B, Sumner, L, Sansone, S, Taylor, C, Hardy, N, and Kaddurah-Daouk, R. Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *OMICS.* 2006:10(2), 158–163.

196. Fiehn, O, Robertson, D, Griffin, J, Van der Werf, M, Nikolau, B, Morrison, N, Sumner, L, Goodacre, R, Hardy, N, Taylor, C, Fostel, J, Kristal, B, Kaddurah-Daouk, R, Mendes, P, van Ommen, B, Lindon, J, and Sansone, S. The metabolomics standards initiative (MSI). *Metabolomics* 2007:3, 175–178.
197. Sumner, L, Amberg, A, Barrett, D, Beale, M, Beger, R, Daykin, C, Fan, T, Fiehn, O, Goodacre, R, Griffin, J, Hankemeier, T, Hardy, N, Harnly, J, Higashi, R, Kopka, J, Lane, A, Lindon, J, Marriott, P, Nicholls, A, Reily, M, Thaden, J, and Viant, M. Proposed minimum reporting standards for chemical analysis. *Metabolomics* 2007:3, 211–221.
198. Searls, D. Data integration: challenges for drug discovery. *Nature Reviews Drug Discovery* 2005:4, 45–58.
199. Redestig, H, Fukushima, A, Stenlund, H, Moritz, T, Arita, M, Saito, K, and Kusano, M. Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Anal.Chem.* 2009:81(19), 7974–7980.
200. Frazier, P, Tix, A, and Barron, K. Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology* 2004: 51, 115–134.
201. Scriver, C. Homeostasis, complexity, and monogenic phenotypes: the view from phenylketonuria. In Valle, D, Beaudet, A, Vogelstein, B, Kinzler, K, Antonarakis, S, Ballabio, A, Scriver, C, Sly, W, Childs, B, Bunz, F, Gibson, K, and Mitchell, G, editors, *Scriver's OMMBID*, 77S1. McGraw-Hill, New York, NY: 2010.

Samenvatting

Dit proefschrift beschrijft de resultaten van verschillende analyses die uitgevoerd kunnen worden in het kader van tweelingenstudies op basis van *metabolomics* data. De tweelingenstudie is een gevestigde methode om te schatten of verschillen tussen personen in meetbare eigenschappen hoofdzakelijk toe te schrijven zijn aan genetische invloeden, danwel aan verschillen in omgeving. *Metabolomics* is een betrekkelijk jonge tak binnen de "*omics*" wetenschappen, die tot doel heeft een uitputtend overzicht te geven van de stoffen (metabolieten) die betrokken zijn bij biochemische processen in biologische systemen. Een onderdeel van *metabolomics* is het meten van de concentraties of onderlinge verhoudingen in concentraties van deze metabolieten in lichaamsvloeistoffen zoals bloed en urine. Centraal in dit proefschrift staan de ontwikkeling en de toepassing van methodes voor analyse van de data die voortkomen uit dergelijke metingen in het kader van (tweelingen)familiestudies. Zodoende draagt dit proefschrift bij aan het ophelderen van de bijdrage van genetische en omgevingsinvloeden aan individuele verschillen in metabolietconcentraties in lichaamsvloeistoffen.

In Hoofdstuk 1 van dit proefschrift wordt een algemene inleiding gegeven in *metabolomics* en tweelingen- en familiestudies. Uiteengezet wordt welke rol familiestudies, en in het bijzonder studies van tweelingen en hun naaste familieleden, hebben voor het bestuderen van de factoren die ten grondslag liggen aan de individuele verschillen voor meetbare eigenschappen die in hun waarde geleidelijk variëren tussen personen. Vanwege hun belangrijke rol in dit proefschrift worden twee methodes geïntroduceerd die kunnen worden gebruikt voor statistische analyse binnen dergelijke studies. De eerste van deze technieken, *structural equation modeling* (SEM), gaat uit van een model dat gebaseerd is op een hypothese met betrekking tot de oorzaken van variatie binnen en tussen verschillende meetbare eigenschappen. In een dergelijk model zijn de bijdragen van de verschillende oorzaken van meetbare variatie opgenomen als parameters

die vrij in waarde kunnen variëren. De parameterwaarden die het beste bij de meetgegevens aansluiten, kunnen worden aangenomen als schattingen voor de waarden van de betreffende parameters in de onderzochte populatiesteekproef.

De tweede techniek voor het bestuderen van de onderlinge verschillen in meetbare eigenschappen die besproken wordt in Hoofdstuk 1 is hiërarchische clusteranalyse. Met behulp van deze techniek kan een overzicht worden verkregen van de onderlinge overeenkomsten tussen variabelen (bijvoorbeeld, metabolieten) of tussen objecten (bijvoorbeeld, proefpersonen) op basis van meetgegevens voor meerdere eigenschappen gemeten in een steekproef. In Hoofdstuk 1 wordt uiteengezet dat in dit proefschrift deze techniek op twee verschillende manieren gebruikt wordt om inzicht te geven in de genetische factoren die ten grondslag liggen aan onderlinge verschillen in meetbare eigenschappen. Voorts wordt in Hoofdstuk 1 een perspectief geschetst hoe studies zoals beschreven in dit proefschrift een overzicht kunnen geven van de genetische en omgevingsinvloeden op de concentraties van verschillende elementen van biologische systemen (bijvoorbeeld, gentranscripten, enzymen en metabolieten) afzonderlijk en in hun samenhang. Tot slot wordt in Hoofdstuk 1 betoogd dat studies op basis van meetgegevens verkregen in bijvoorbeeld tweelingenfamilies, de interpretatie van genoom-brede associatiestudies kunnen verbeteren en onder andere daarmee een bijdrage kunnen leveren aan de opheldering van zogenaamde complexe aandoeningen.

Hoofdstuk 2 beschrijft de onderlinge verschillen en overeenkomsten in lipidenprofielen zoals gemeten in bloedplasma tussen (voornamelijk ééneiige) tweelingen en hun niet-tweelingbroers en -zussen. Deze lipidenprofielen werden verkregen met één van de meest gebruikte meetmethodes binnen *metabolomics*, te weten vloeistofchromatografie gekoppeld aan massaspectrometrie (LC–MS). Bij deze techniek worden de componenten in het onderzochte monster eerst gescheiden op een chromatografische kolom op basis van hun verschillen in fysisch-chemische eigenschappen, en vervolgens gedetecteerd met een massaspectrometer. In de studie zoals beschreven in Hoofdstuk 2 werden in het bloedplasmamonster van iedere proefpersoon met LC–MS relatieve concentraties gemeten van in totaal 61 verschillende lipiden uit vijf verschillende lipidenklassen. De gemeten lipiden zijn betrokken bij een breed scala van fysiologische en pathofysiologische processen, waaronder signaaltransductie, ontstekingsreacties en energiehuishouding.

Hiërarchische clusteranalyse werd in Hoofdstuk 2 gebruikt om de proefpersonen te groeperen op basis van hun onderlinge verschillen en overeenkomsten in plasmalipidenprofiel. Het resultaat van deze groepering suggereerde dat mannelijke en vrouwelijke proefpersonen verschillende lipidenprofielen hadden. Verdere analyses van de resultaten van de hiërarchische clusteranalyse onderbouwden de hypothese dat in het algemeen, overeenkomsten in genetische en omgevingsinvloeden bijdragen aan overeenkomsten in lipidenprofielen tussen personen. Echter, in het onderzoek zoals beschreven in Hoofdstuk 2 werden ook aanwijzingen gevonden dat bepaalde omstandigheden, waaronder recente ziekte, samengaan met veranderingen in het lipidenprofiel zoals gemeten in

bloedplasma.

Het onderscheidend vermogen van statistische toetsen wordt vergroot door het aantal waarnemingen te vergroten op basis waarvan getoetst wordt. In dit verband beschrijft Hoofdstuk 3, een techniek genaamd "*quantile equating*", die het mogelijk maakt meetgegevens te combineren die verkregen zijn met dezelfde semi-kwantitatieve analytisch-chemische methode, maar bijvoorbeeld op verschillende tijdstippen binnen een grote studie. In dit hoofdstuk wordt beargumenteerd dat het gebruik van dergelijke datavoorbewerkingstechnieken noodzakelijk kan zijn vanwege praktisch onvermijdbare kleine verschillen in analytische methodologie tussen 'blokken' van metingen. De succesvolle toepassing van de in dit hoofdstuk gentroduceerde methode wordt gedemonstreerd aan de hand van meetgegevens verkregen met LC–MS metingen van relatieve concentraties van lipiden in bloedplasma, en met metingen van protonenkernspinresonantie (^{1}H NMR) in bloedplasma en in urinemonsters. Alle waterstofhoudende moleculen in een monster dragen bij aan het met ^{1}H NMR gedetecteerde signaal, en daarmee bevatten ^{1}H NMR data informatie over de concentraties van metabolieten behorend tot verschillende klassen.

De met behulp van de in Hoofdstuk 3 beschreven methode gecombineerde LC–MS data, werden vervolgens gebruikt voor de studie zoals beschreven in Hoofdstuk 4. Deze samengestelde dataset bevatte meetgegevens voor ééneiige tweelingen en hun niet tweelingbroers en -zussen zoals beschreven in Hoofdstuk 2, en gegevens voor twee-eiige tweelingen en hun niet tweelingbroers en -zussen. Evenals in Hoofdstuk 2 werd hiërarchische clusteranalyse toegepast om de onderlinge variatie in lipidenprofielen tussen personen in kaart te brengen. De aanwezigheid van twee-eiige tweelingen in deze studie, maakte het mogelijk om de invloed van gedeelde omgevingsinvloeden op overeenkomsten in deze profielen beter vast te kunnen stellen. Tevens werden, evenals in Hoofdstuk 2, aanwijzingen gevonden dat bijvoorbeeld ziekte in het recente verleden samen zou kunnen gaan met veranderingen in het lipidenprofiel in bloedplasma. Daarnaast bevestigden de resultaten van de hiërarchische clusteranalyse dat toepassing van een methode zoals beschreven in Hoofdstuk 3 noodzakelijk was geweest om de meetgegevens verkregen in verschillende blokken samen te kunnen voegen.

Deze zelfde gecombineerde meetgegevens op basis van LC–MS analyse van lipiden in bloedplasma, evenals de samengevoegde gegevens verkregen met ^{1}H NMR metingen in bloedplasma, werden gebruikt voor de analyses zoals beschreven in Hoofdstuk 5. Allereerst werden in dit onderzoek met SEM de relatieve bijdragen geschat van genetische invloeden en van omgevingsinvloeden aan de verschillen tussen personen in de concentraties gemeten voor elke afzonderlijke variabele. Dergelijke analyses van de triglyceriden gemeten met LC–MS lieten een consistent patroon zien, waarbij zowel het aantal koolstofatomen als het aantal dubbele bindingen in de vetzuurstaarten van belang leek voor de erfelijkheid van de gemeten concentraties.

Vervolgens werden in een bivariate SEM-analyse, voor elke paarsgewijze combinatie van variabelen gemeten met elk van beide analytische methoden,

de genetische correlaties tussen variabelen berekend. Hiërarchische clusteranalyse werd gebruikt om de patronen in deze genetische correlaties voor zowel de LC–MS gegevens als de ^{1}H NMR gegevens te onderzoeken. Hierbij viel op dat de positieve correlatie tussen verschillende lipiden in bloedplasma behorend tot dezelfde lipidenklasse, samen leek te hangen met gedeelde erfelijke invloeden tussen deze lipiden. Binnen de variabelen gemeten met ^{1}H NMR werden grote verschillen in de genetische correlaties waargenomen, die te maken zouden kunnen hebben met het feit dat met de gebruikte ^{1}H NMR methode metabolieten van verschillende metabolietklassen waargenomen kunnen worden.

Enkele conclusies en aanbevelingen voor verder onderzoek naar aanleiding van dit proefschrift worden beschreven in Hoofdstuk 6. Één van de conclusies is dat methodes toegepast binnen *microarray* onderzoek voor het combineren van data gemeten voor dezelfde gentranscripten maar in verschillende personen, mogelijk toepasbaar zijn binnen *metabolomics* voor het combineren van data gemeten voor dezelfde metabolieten maar in verschillende personen. Deze methodes kunnen daarmee een aanvulling vormen op de methode zoals beschreven in Hoofdstuk 3 van dit proefschrift. Ook wordt in Hoofdstuk 6 betoogd dat toepassing van de relatief 'hypothese-vrije' methode voor clustering van proefpersonen zoals beschreven in Hoofdstukken 2 en 4 resultaten gaf die consistent waren met een vooraf opgestelde hypothese. Anderzijds werd de 'hypothese-gedreven' methode SEM in Hoofdstuk 5 op een relatief hypothese-vrije manier toegepast, wat evenwel eveneens resultaten gaf die consistent waren met de biologische achtergrond van de data. Een mogelijke toepassing van de resultaten van hiërarchische clusteranalyse van proefpersonen voor het vinden van voor SEM relevante covariaten wordt beschreven.

Tot slot wordt aangegeven dat methodes om ziekte op te sporen door de analyse van afstanden tussen personen, zoals bijvoorbeeld beschreven in Hoofdstukken 2 en 4 van dit proefschrift, verder onderzoek verdienen vanwege het mogelijk hogere onderscheidend vermogen ten opzichte van methodes gebaseerd op analyse van afzonderlijke variabelen.

Curriculum Vitae

The author of this thesis was born in 1981 in Voorburg, The Netherlands. In 1993 he was admitted to the St.-Maartenscollege in that same place, where he completed his pre-university secondary education (VWO) in 1999. In that same year, he started his studies Biomedical Sciences at the Leiden University Medical Center (LUMC) (Leiden, The Netherlands). He conducted his first internship at the LUMC Department of Nephrology under the guidance of dr. S.A. Joosten, where he studied the immune reaction against mesangial cells in the kidney following kidney transplantation. His second internship was at the Department of Cardiology of the LUMC with dr. C.A. Swenne, where the author investigated electrocardiographic methods to assess repolarization heterogeneity in the heart and developed the research-oriented ECG analysis software "LEADS". After his graduation in 2005, he worked for a short period at this same department as a scientific research assistant. The author started his PhD studies at the Division of Analytical Biosciences of the Leiden/Amsterdam Center for Drug Research (LACDR) (Leiden, The Netherlands) in February 2006. Since December 2010 he works as a postdoctoral researcher at the department of Biological Psychology of the VU University in Amsterdam, The Netherlands.

List of Publications

Draisma, HHM, Reijmers, TH, Bobeldijk-Pastorova, I, Meulman, JJ, Estourgie-Van Burk, G, Bartels, M, Ramaker, R, van der Greef, J, Boomsma, DI, and Hankemeier, T. Similarities and differences in lipidomics profiles among healthy monozygotic twin pairs. *OMICS* 2008:12(1), 17–31

Draisma, HHM, Reijmers, TH, van der Kloet, F, Bobeldijk-Pastorova, I, Spies-Faber, E, Vogels, JTWE, Meulman, JJ, Boomsma, DI, Van der Greef, J, and Hankemeier, T. Equating, or correction for between-block effects with application to body fluid LC–MS and NMR metabolomics data sets. *Anal.Chem.* 2010:82(3), 1039–1046

Draisma, HHM, Reijmers, TH, Meulman, JJ, Boomsma, DI, van der Greef, J, and Hankemeier, T. Hierarchical clustering analysis of blood plasma lipidomics profiles from mono- and dizygotic twin families. *In preparation*

Draisma, HHM, Reijmers, TH, Meulman, JJ, Boomsma, DI, van der Greef, J, and Hankemeier, T. Contribution of genetic and environmental factors to variation in the human metabolome: a multivariate study in twins and siblings. *Submitted for publication*

Not in this thesis:

van Huysduynen, BH, Swenne, CA, Bax, JJ, Bleeker, GB, Draisma, HHM, van Erven, L, Molhoek, SG, van de Vooren, H, van der Wall, EE, and Schalij, MJ. Dispersion of repolarization in cardiac resynchronization therapy. *Heart Rhythm* 2005:2(12), 1286–1293

van Huysduynen, BH, Swenne, CA, Draisma, HHM, Antoni, ML, Vooren, HVD, van der Wall, EE, and Schalij, MJ. Validation of ECG indices of ventricular repolarization heterogeneity: a computer simulation study. *J.Cardiovasc.Electrophysiol.* 2005:16(10), 1097–1103

Draisma, HHM, Schalij, MJ, van der Wall, EE, and Swenne, CA. Elucidation of the spatial ventricular gradient and its link with dispersion of repolarization. *Heart Rhythm* 2006:3(9), 1092–1099

Henkens, IR, Mouchaers, KTB, Vliegen, HW, van der Laarse, WJ, Swenne, CA, Maan, AC, Draisma, HHM, Schalij, I, van der Wall, EE, Schalij, MJ, and Vonk-Noordegraaf, A. Early changes in rat hearts with developing pulmonary arterial hypertension can be detected with three-dimensional electrocardiography. *Am.J.Physiol.Heart.Circ.Physiol.* 2007:293(2), H1300–H1307

Swenne, CA, van Huysduynen, BH, Bax, JJ, Bleeker, GB, Draisma, HHM, van Erven, L, Molhoek, SG, van de Vooren, H, van der Wall, EE, and Schalij, MJ. Biventricular pacing and transmural dispersion of the repolarization. *Europace* 2007:9(1), 48–49

van Huysduynen, BH, Henkens, IR, Swenne, CA, Oosterhof, T, Draisma, HHM, Maan, AC, Hazekamp, MG, de Roos, A, Schalij, M, van der Wall, EE, and Vliegen, HW. Pulmonary valve replacement in tetralogy of Fallot improves the repolarization. *Int.J.Cardiol.* 2008:124, 301–306

Man, S, Maan, AC, Kim, E, Draisma, HHM, Schalij, MJ, van der Wall, EE, and Swenne, CA. Reconstruction of standard 12-lead electrocardiograms from 12-lead electrocardiograms recorded with the Mason-Likar electrode configuration. *J.Electrocardiol.* 2008:41, 211–219

Scherptong, RWC, Henkens, IR, Man, SC, Cessie, SL, Vliegen, HW, Draisma, HHM, Maan, AC, Schalij, MJ, and Swenne, CA. Normal limits of the spatial QRS-T angle and ventricular gradient in 12-lead electrocardiograms of young adults: dependence on sex and heart rate. *J.Electrocardiol.* 2008:41(6), 648–655

Nawoord

Alles verandert. Dit nawoord biedt de gelegenheid om een aantal mensen te bedanken die hebben bijgedragen aan veranderingen waarvan het wetenschappelijke resultaat beschreven staat in de rest van dit proefschrift. Op deze plaats betuig ik dan ook mijn dank aan degenen die mij de mogelijkheid hebben gegeven mijzelf te ontwikkelen, wetenschappelijk of anderszins. Deze groep mensen is te groot en hun bijdragen te divers om iedereen hier persoonlijk te noemen; dit betekent echter niet dat ik hen die hier niet vermeld zijn geen dank verschuldigd ben.

De afdeling Analytische Biowetenschappen in Leiden is zeker in de afgelopen jaren te groot geworden om iedereen te bedanken, maar een aantal mensen met wie ik bijzonder veel lief en leed gedeeld heb noem ik hier in het bijzonder. Jurre, sinds we op dezelfde kamer zaten heb ik je aanstekelijke discipline van dichtbij mogen meemaken. Peter en Jan-Willem, jullie waren misschien wel de AIOs die voor het meeste leven in de brouwerij zorgden; we kunnen in ieder geval trots zijn op ons cinematografische hoogstandje... Loes, ik ben blij dat je ondanks mijn grillen altijd bereid was om iets te regelen of uit te zoeken als dat nodig was. Shanna bedankt voor alle vrolijkheid en gezelligheid die jou omringt. Kjeld, de situaties waarin we elkaar de afgelopen jaren tegenkwamen kunnen in ieder geval als bijzonder bestempeld worden... ("ga je nog iets doen vanavond?"). Ubbo, ondanks alle veranderingen bedank ik ook jou voor je bijdrage aan mijn ontwikkeling. Robert, dank voor je praktische tips in de eindfase van mijn promotie.

Ook de medewerkers van de Gorlaeus helpdesk wil ik bedanken voor al hun hulp.

Beste Theo, bedankt voor je feedback, je praktische benadering, en voor alle goede gesprekken die mij zeker nieuwe moed hebben gegeven als het allemaal wat minder ging. Ik prijs me gelukkig je als stabiele factor aan mijn zijde te hebben gehad. Dorret, ik ben je zeer erkentelijk voor je duidelijk zichtbare

bijdrage aan de wetenschappelijke inhoud van dit proefschrift. Tevens ben ik je dankbaar dat ik de afdeling Biologische Psychologie steeds beter kan leren kennen.

Steven, het leven is weliswaar te kort voor vrienden, maar je vormt wat mij betreft zelf een grote uitzondering op die regel. Bram, Mark en Erik, ik ben benieuwd in welk oord het volgende clanweekend gaat plaatsvinden. CIA, bedankt dat jullie me over mijn angst voor Laven hebben geholpen. Janneke, dank dat je me daarnaast ook nog hebt bijgestaan met nuttige adviezen. Kun en Guido, laten we binnenkort weer eens een leuk restaurantje uitzoeken! Inge, dank voor je geduld tijdens de vele lessen waarin je al geprobeerd hebt van mij een beter musicus te maken. Yvonne, bedankt voor alle tijd die ik met je heb mogen doorbrengen. En Gulle Herbergiers, na meer dan vijf jaar heb ik eindelijk de kans om de rekening te vereffenen: bedankt voor alle morele en andere ondersteuning die jullie me altijd hebben gegeven. Ik hoop dat jullie nog lang mijn veilige haven kunnen zijn.