Universiteit
Leiden
The Netherlands

Analysis of metabolomics data from twin families

Draisma, H.H.M.

Citation
Draisma, H. H. M. (2011, May 10). *Analysis of metabolomics data from twin families*. Retrieved from https://hdl.handle.net/1887/17643

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/17643

Note: To cite this publication please use the final published version (if applicable).

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

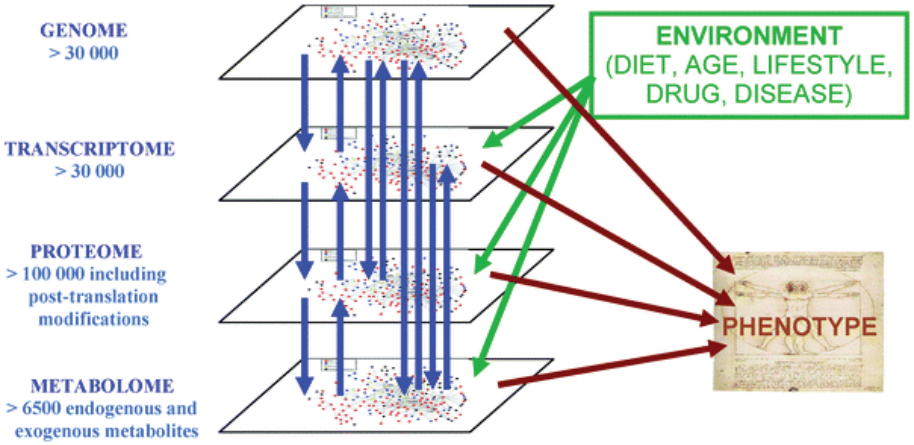

Figure 1.1: The different information levels within (human) cells and their interactions. The phenotype is a function of the interactions of the components of genome, transcriptome, proteome, and metabolome with each other and with the environment. According to this view, the environment might exert 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

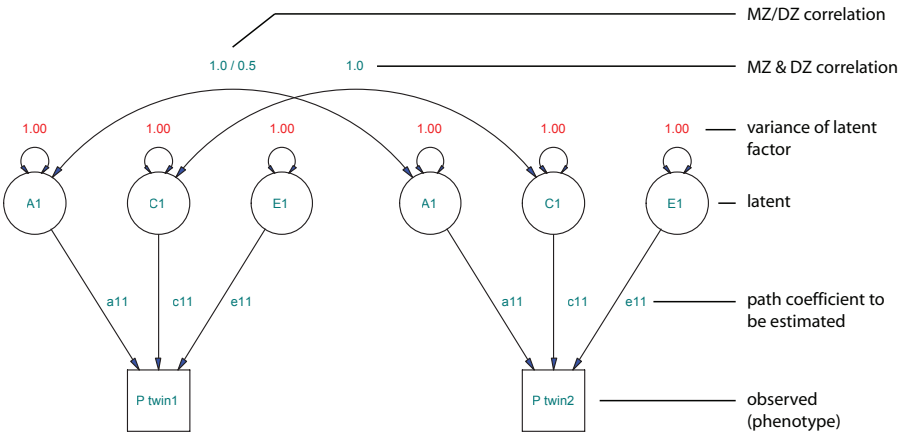

Figure 1.2: Path diagram for univariate quantitative genetic analysis under classical twin design. Latent (unobserved) factors and path coefficients are indicated by uppercase (*e.g.*, "A") and lowercase (*e.g.*, "a") letters, respectively. In a path coefficients model, the variances of the factors are 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

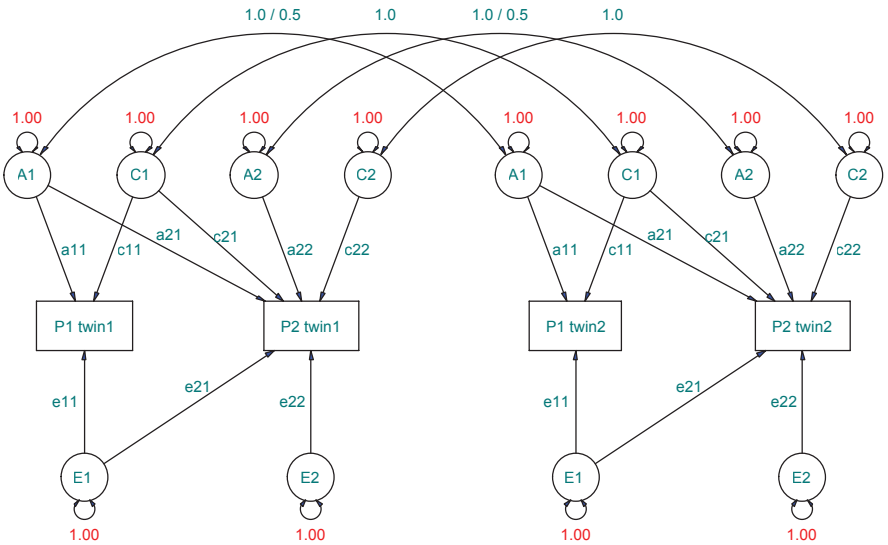

Figure 1.4: Path diagram for bivariate quantitative genetic analysis under classical twin design. A1 and A2 are additive genetic factors, where A1 is possibly shared by both traits and A2 is specific for phenotype 2, respectively; similarly for C1 and C2 (familial environment), and for E1 and E2 (specific 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

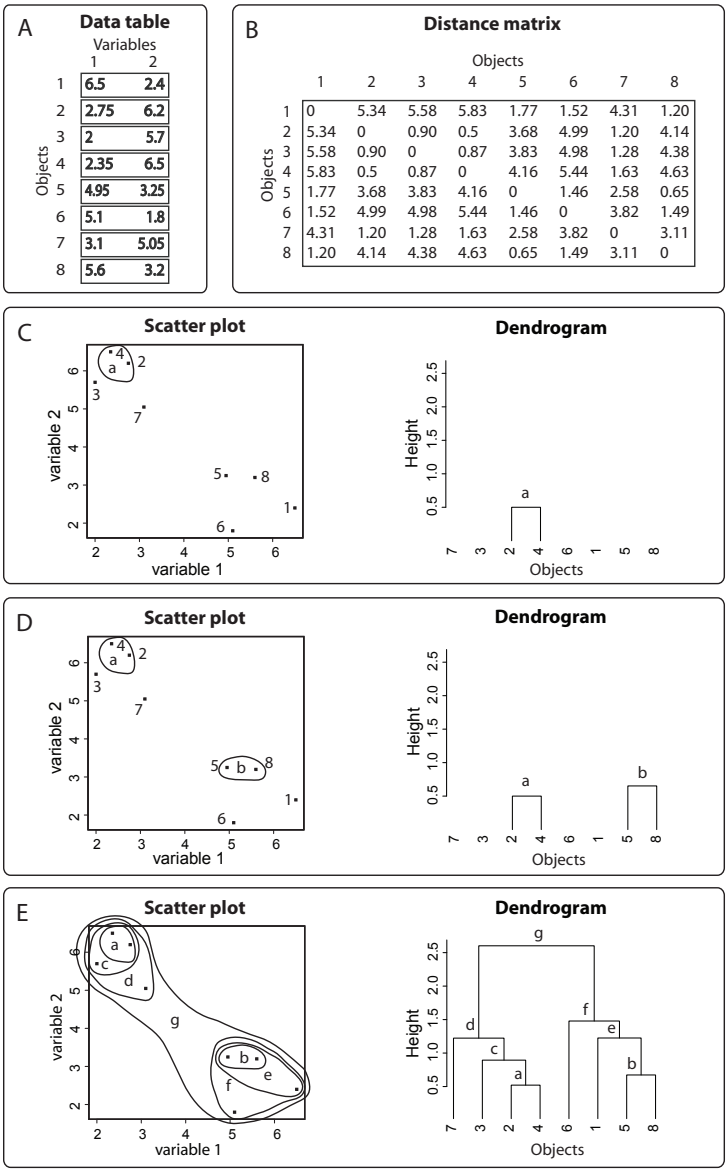

Figure 1.6: Example of hierarchical clustering analysis of eight objects in a bivariate data matrix. The "height" of the branching points in the dendrogram (C–E) equals the distances among clusters as computed by the hierarchical clustering algorithm on the basis of the distance matrix (B). These branching 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-

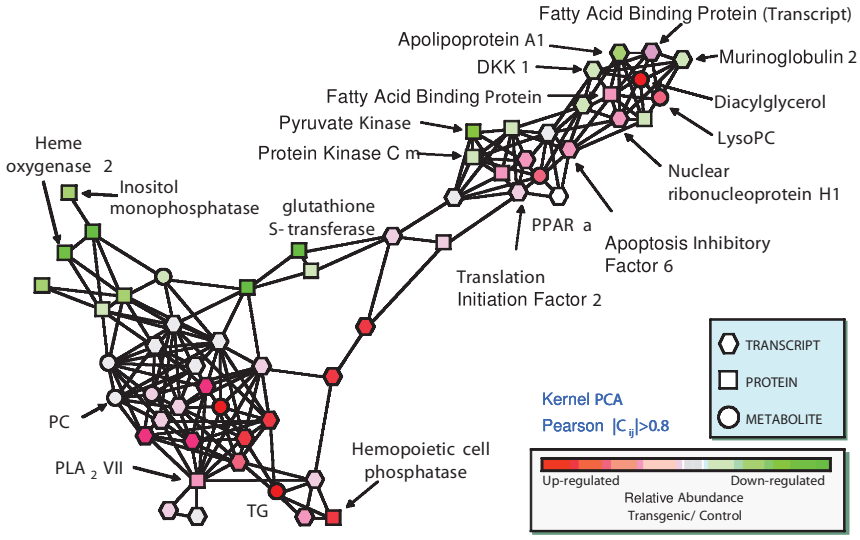

Figure 1.7: Correlation network of select genes, proteins and lipids in the APOE*3 Leiden mouse. The colors of the elements of the system (transcripts, proteins and metabolites) indicate the relative expression in the transgenic animals compared to wild type controls (red=higher level; green=lower level) 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.