Analysis of metabolomics data from twin families

Draisma, H.H.M.

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

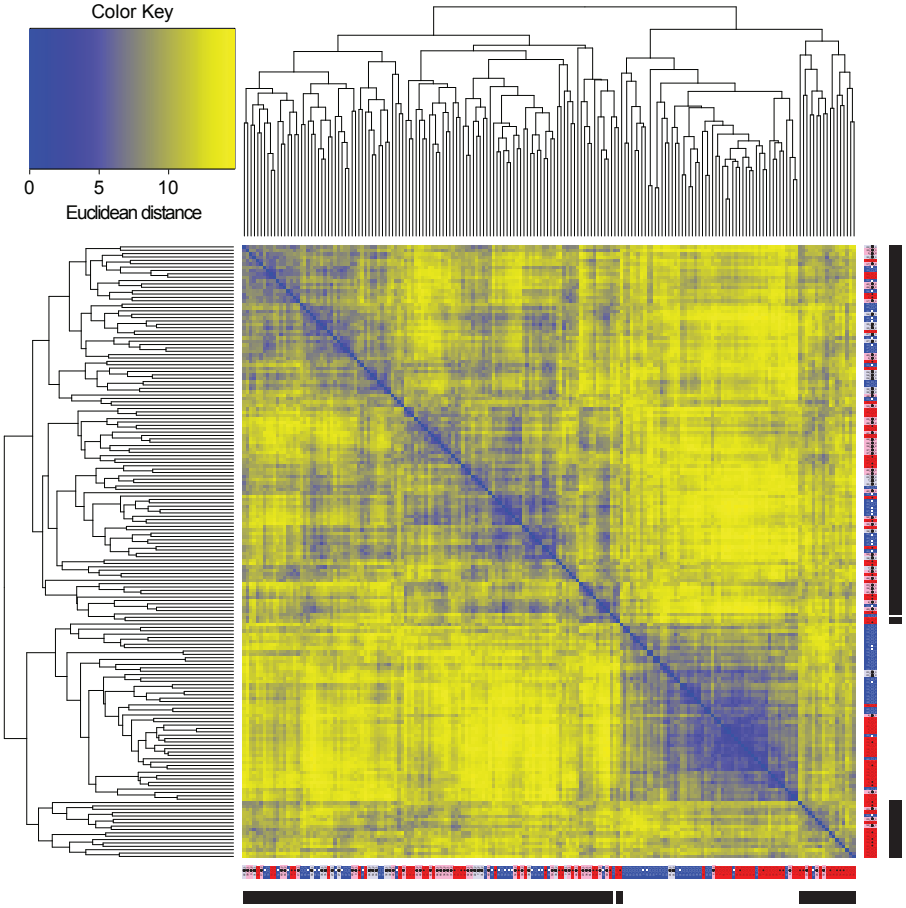

Figure 4.1: Heatmap of Euclidean distances between objects, and associated hierarchical clustering dendrograms for the combined (concatenated with variables as shared mode) B1–B2 data set before quantile equating. In this figure, individual objects are labeled by two color codes: the first color encodes 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.

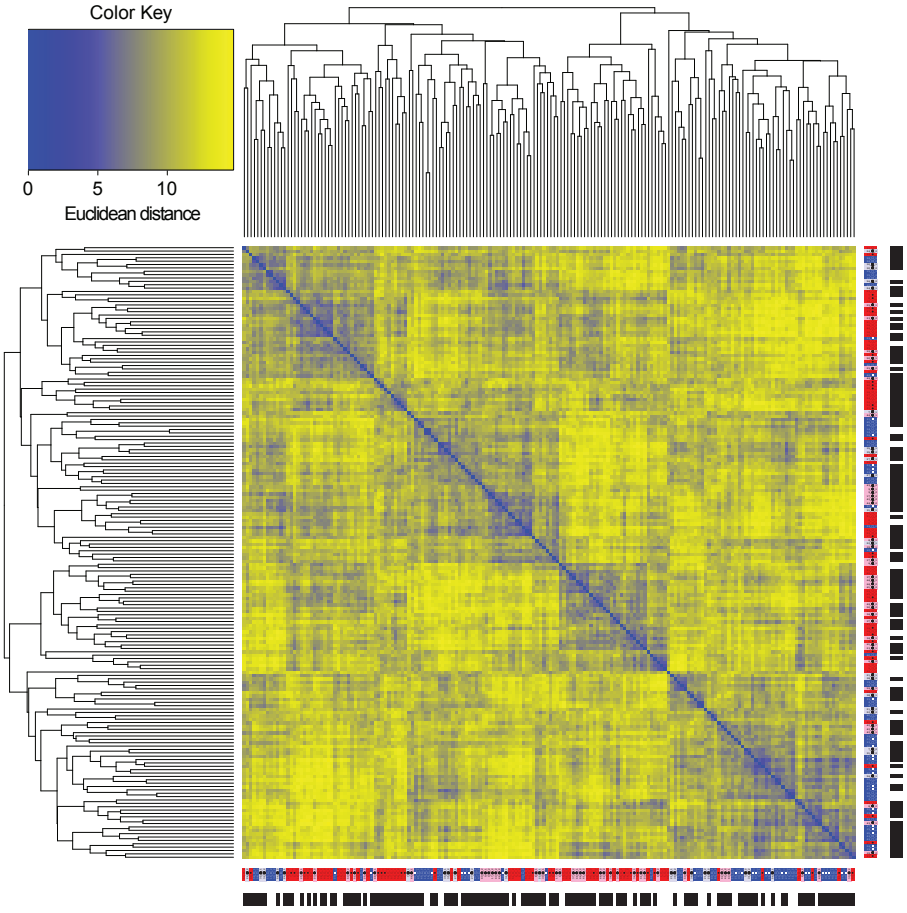

Figure 4.2: Heatmap of Euclidean distances between objects, and associated hierarchical clustering dendrograms for the combined (concatenated with variables as shared mode) B1–B2 data set after quantile equating. For legend, see Figure 4.1.

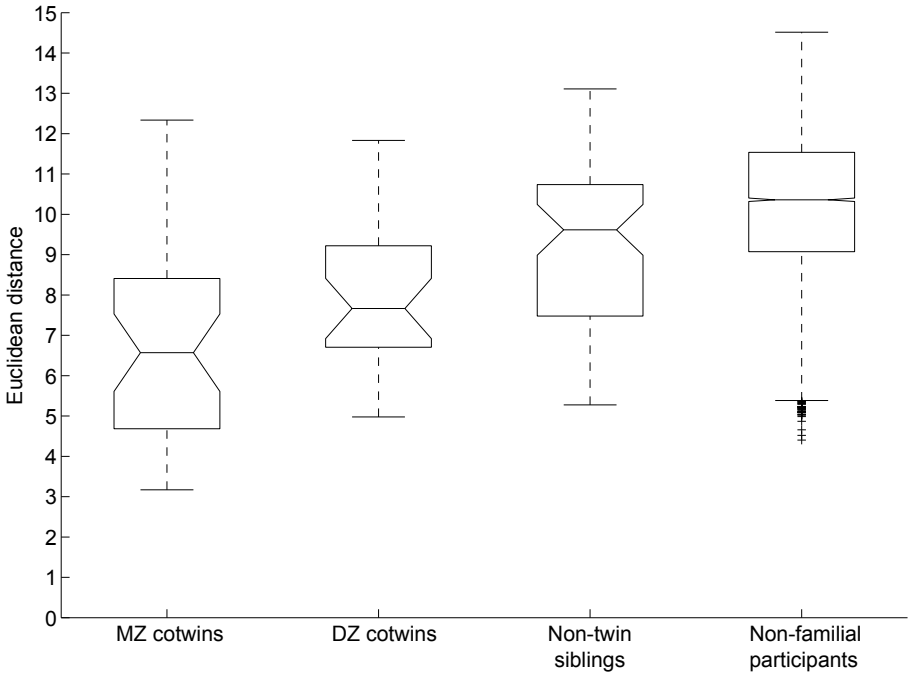

Figure 4.3: Box-whisker plots showing distributions of Euclidean distances between MZ co-twins (N=37), between DZ co-twins (N=28), among sex-matched nontwin siblings (N=66), and among sex-matched nonfamilial participants (N=8,203) in the combined equated B1–B2 data set. The observations indicated with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

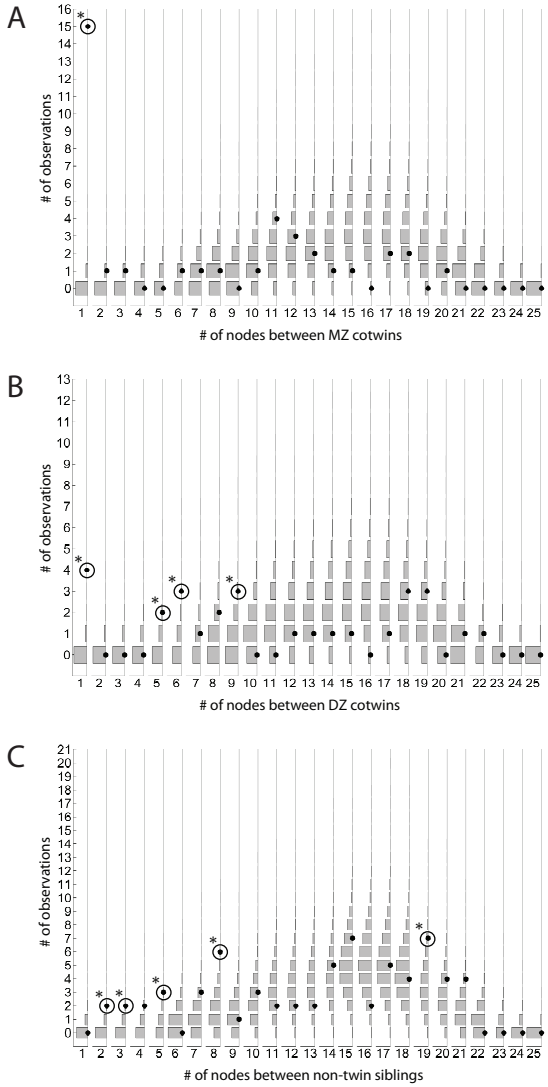

Figure 4.4: Results of node analyses for MZ co-twins (A), DZ co-twins (B), and sex-matched nontwin siblings (C) with respect to permutation-based chance distributions. Numbers of nodes separating co-twins or nontwin siblings increase from left to right in each panel. For each number of branching points, 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

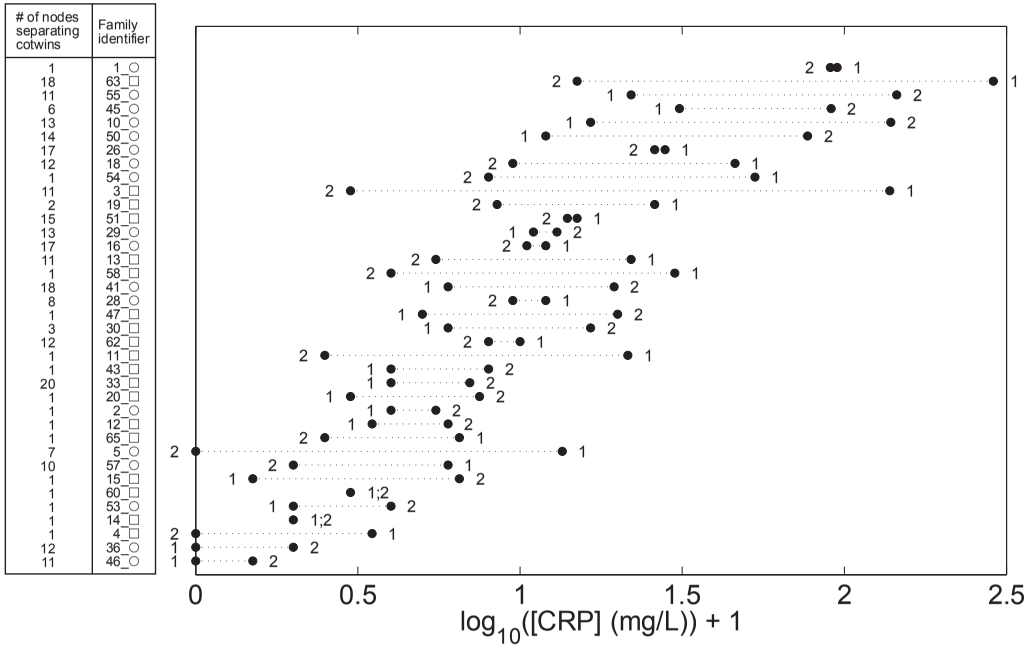

Figure 4.6: C-reactive protein (CRP) levels in blood samples from MZ twins. From bottom to top, the average CRP level of twin pairs increases. On the left hand side, for each MZ twin pair the number of nodes separating co-twins in the dendrogram is displayed. The numbers "1" and "2" near the observations 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.