# Similarities and Differences in Lipidomics Profiles among Healthy Monozygotic Twin Pairs

**HARMEN H.M. DRAISMA,[1] THEO H. REIJMERS,[1]
IVANA BOBELDIJK-PASTOROVA,[2] JACQUELINE J. MEULMAN,[3]
G. FREDERIEK ESTOURGIE-VAN BURK,[4,5] MEIKE BARTELS,[5]
RAYMOND RAMAKER,[2] JAN VAN DER GREEF,[1] DORRET I. BOOMSMA,[5]
and THOMAS HANKEMEIER[1]**

## ABSTRACT

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

[1]Leiden University, LACDR, Leiden, The Netherlands.

[2]TNO Quality of Life, Zeist, The Netherlands.

[3]Leiden University, Mathematical Institute, Leiden, The Netherlands.

[4]Department of Paediatric Endocrinology, Institute for Clinical and Experimental Neurosciences, VU University Medical Center, Amsterdam, The Netherlands.

[5]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands.

## INTRODUCTION

**D**IFFERENCES IN GENETIC MAKEUP and in environmental exposure manifest as differences in measurable characteristics in individuals, that is, as differences in phenotypes (Fischer et al., 2004). Metabolite profiles are regarded as being an important part of the phenotype (Goodacre, 2007). It is currently unknown to what extent an individual's metabolite profile is a function of the genotype and of environmental conditions. If genotype is an important determinant of metabolite profiles, it is expected that biological relatives who share genes and possibly also share environments will show similarities in metabolic profiles. To explore these issues, we carried out a study in healthy, 18-year-old monozygotic (MZ) twins and their biological siblings. The process of obtaining a comprehensive view of the metabolites in an organism has been termed "metabolomics" (Goodacre, 2007). In humans, metabolomics strategies are often used to find differences in metabolite profiles between groups having different phenotypes, for example, between groups of healthy and diseased individuals (Van der Greef et al., 2004). Indeed, with respect to other 'omes such as the genome, the metabolome might be more informative of the physiological state of an organism. For example, in a study where similarities in gene expression profiles of twins discordant for rheumatoid arthritis were compared to similarities in expression between healthy twins, no difference was found between the healthy twin pairs and the twin pairs where one twin had the disease (Nanki et al., 1996). Perhaps the most widely used techniques to measure a wide range of metabolites in biological samples in metabolomics are nuclear magnetic resonance (NMR) and gas or liquid chromatography coupled to mass spectrometry (GC-MS and LC-MS, respectively). NMR aims at obtaining a picture of the complete metabolite profile of a sample, and thus is able to provide a "global" view of the metabolome. Its sensitivity is typically lower than that of MS-based methods such as LC-MS, though. A "targeted" approach, on the other hand, focuses on analysis of particular classes of metabolites, for example amino acids, sterols, or lipids (German et al., 2003). An LC-MS platform can be used for both global and targeted approaches (Van der Greef et al., 2004), but it is impossible to analyze in one run metabolites that have widely differing physicochemical properties such as different polarities and acid dissociation constants. Disadvantages of gas chromatography when applied in metabolomics studies are that often derivatization is necessary (Koek et al., 2006), and that even then, only particular classes of metabolites are measurable. In this study we have applied LC-MS in a targeted manner to obtain lipid profiles in blood plasma samples from healthy MZ and dizygotic (DZ) twin pairs and their siblings. Previous research in our laboratory using the LC-MS method applied in the current study suggested that family members had relatively similar blood plasma lipid profiles, although strong evidence for this was lacking (unpublished results). Furthermore, lipids are especially interesting metabolites because they are involved in a wide range of physiological processes. For example, triglycerides (TGs) serve as an energy source for the body (Murphy, 2001), as a precursor for cell membrane phospholipids (Baur et al., 2000), and in the form of body fat they are important for thermal insulation (Swift, 1932). Among the lipids, TGs are the most important class into which potentially toxic compounds can be incorporated (Dodds, 1991). Another class of lipids with entirely different functions comprises the lysophosphatidylcholines (LPCs). These can be formed from phosphatidylcholines (PCs) present in low-density lipoprotein, for example, by platelet-activating factor acetylhydrolase (lipoprotein-associated phospholipase $A_2$) (Tew et al., 1996). The activity of this enzyme may be increased upon proinflammatory stimuli (Elstad et al., 1989); the formed LPCs can act as a chemoattractant for phagocytes (Quinn et al., 1988). PCs can also act as fatty acid donor for cholesterol esterification by the LCAT enzyme (Glomset, 1968), and may cause platelet aggregation after their oxidation (Kern et al., 1998). Bile is partly comprised of PCs (Coleman, 1987). Furthermore, PCs are precursors of sphingomyelins (SPMs), and share some of their functions with them: lipids from both classes are important structural components of cell membranes (Zachowski, 1993) and of lipoprotein particles (McKeone et al., 1993). They are also involved in signal transduction (Sigal et al., 2005), and are constituents of lung surfactant (Veldhuizen et al., 1998). Whereas the surface of a lipoprotein partly consists of PCs and SPMs, cholesteryl esters (ChEs) are an integral part of its core (McKeone et al., 1993). The main biological function of ChEs is that they are precursors of steroids (Acton et al., 1996). Twins are particularly informative study populations because the members of pairs share genetic and environmental influences. MZ co-twins share their complete or nearly complete DNA sequence.

Thus, for any heritable trait, they will show phenotypic resemblance. The more heritable a trait, that is, the larger the influence of additive genetic variation on the phenotype, the larger the resemblance in MZ twins. First-degree relatives such as DZ twins and biological siblings share on average 50% of their segregating genes. Therefore, also for these relatives their phenotypic resemblance is expected to be considerably dependent upon the heritability of the traits under consideration. However, resemblance between relatives who are not MZ twins also depends on the genetic architecture of a trait. For example, if non-additive genetic influences such as dominance or epistasis are of importance, phenotypic resemblance in siblings is expected to be relatively low. If, on the other hand, genetic influences are mainly additive, phenotypic resemblance in DZ twins and siblings will be roughly half of the resemblance in MZ twins. If, next to heritability, the shared family environment—in the literature also referred to as the "common environment" or "family environment" (Neale and Cardon, 1992)—also contributes to phenotypic resemblance of relatives, then first-degree relatives will approach the resemblance of MZ twins more than is expected on basis of genetic segregation. In classical twin studies, knowledge about genetic and social relationships among co-twins and siblings reared together is used to impose certain structure upon the measurement data (Neale and Cardon, 1992). Uni- and multivariate data are often modeled within the context of genetic covariance structure approaches, using estimation techniques based on maximum likelihood. However, such other approaches require that the number of measured variables is not (much) larger than the number of independent clusters (e.g., twin pairs or families) that take part in the study. Therefore, such techniques have rather limited applicability in typical 'omics studies, where the number of measured characteristics is much larger than the number of individual samples. Such a multi- or megavariate approach is the consequence of the idea that when studying biological systems, multiple rather than individual measured variables will reflect underlying, as such unobserved, phenomena. As an alternative, in the current study we have applied an unsupervised approach that is based upon hierarchical clustering of metabolite profiles to identify biologically relevant subgroups of participants (i.e., twin pairs and families) in the data. With this approach, it is possible to get an impression of the within-family variation in metabolite profiles relative to the between-family variation. We expected to identify clusters of family members in the data, in those cases where family members share relevant genes and/or environment. Coclustering of twins was evaluated using a permutation test. In instances where co-twins did not cluster closely together, we have attempted to provide explanations for this. Our results suggest an important role of genetic background in the generation of interindividual variation in blood plasma lipid profiles. Moreover, several lipids measured in this study may prove to be appropriate for monitoring health status, for example, in the course of personalized treatment.

## METHODS

### Participants

Participants were recruited from the Netherlands Twin Register at the Vrije Universiteit (VU) in Amsterdam, The Netherlands (Boomsma et al., 2006). The aim was to recruit MZ twin pairs of approximately 18 years old from a cohort participating in a longitudinal investigation into the heritability of mental and physical development in late puberty (Hoekstra et al., 2007). Near the twins' 18th birthday, the twin pairs and their siblings were invited to take part in the project. Ethical approval was given by the Central Committee on Research Involving Human Subjects in The Netherlands. Informed consent and parental consent, if a sibling was under 18, were obtained. Zygosity was determined for all twin pairs by DNA genotyping ($N = 20$ pairs) or using blood group polymorphisms ($N = 1$ pair). Between November 2004 and September 2005, all participants came to the VU University in Amsterdam for a physical examination in the morning and neurophysiological assessment in the afternoon. Blood was drawn after overnight fasting during the morning session. In addition, subjects completed a series of questionnaires regarding demographics, problem behavior, health, lifestyle, educational attainment, and other traits. For the current study, we used answers to questions regarding current use of any medication, subjective health up to 1 month prior to blood sampling, current and earlier smoking habits, and whether participants currently lived at their parents' home.

*Blood sampling*

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

*Sample preparation*

From each plasma sample, 10 $\mu$L aliquots were taken in duplicate. For quality control purposes a pooled sample consisting of equal amounts of plasma from all study participants was prepared and divided into 10 $\mu$L aliquots. These samples (QC samples) were further treated in the same way as the study samples. The samples were divided into two batches, each batch containing one aliquot of each study sample. After separate randomization of each batch QC samples were inserted following each ninth study sample. Samples were deproteinized by adding 300 $\mu$L of isopropanol containing the following internal standards: C17:0 LPC 1 $\mu$g/mL, C24:0 PC 1 $\mu$g/mL, C17:0 ChE 1 $\mu$g/mL, and C51:0 TG 1$\mu$g/mL. In this denomination of lipids, the number of carbon atoms as well as the number of double bonds in the fatty acid, separated by a colon (e.g., C17:0) are followed by the class abbreviation (e.g., LPC). After centrifugation, the clear supernatant was collected and the samples were again stored at $-20°C$ until analysis.

*LC-MS lipid profiling*

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

*Data processing/integration*

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

*Assessment of the quality of the data*

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

*Statistical analysis*

Statistical analyses were carried out in the statistical language and environment R (version 2.2.1) (R Development Core Team, 2005) and in MATLAB (version R2006b, The Mathworks, Natick, MA). For each

sample the replicate measurements were averaged. The resulting data matrix was autoscaled, rendering the mean of the distribution for each lipid compound zero and its variance around this mean one, with the aim to assign all lipid compounds equal weight in the subsequent hierarchical clustering (Vandeginste et al., 1998). Then, each row of the data matrix, corresponding to the averaged profile of one study participant and henceforth denoted as "object", was subjected to standard normal variate scaling (SNV) (Barnes et al., 1989; Sokal, 1961) to correct for the interindividual differences in the total lipid signal observed by this method. Euclidean distances were computed to measure the dissimilarities among objects. According to the Young–Householder theorem, SNV (applied to the objects) followed by squared Euclidean distance computation is mathematically equivalent to computing $(1-)$ the correlation among unscaled objects (Young and Householder, 1938). To assess whether there were differences in median Euclidean distance among (1) MZ co-twins, (2) MZ twins and their same-sex siblings, and (3) same-sex nonfamilial study participants, we performed a multiple comparison procedure using a Tukey's honestly significant difference criterion type of critical value on basis of the result of a nonparametric analysis of the variance within these groups of study participants versus the variance between groups (Kruskal and Wallis, 1952). A multiple comparison procedure is designed to be conservative when testing for significant differences among pairs of groups (Hochberg and Tamhane, 1987). Subsequently, the calculated distances among all objects were subjected to hierarchical clustering analysis. In our choice of the used clustering algorithm we strived for maximum correlation of the distances among clusters as computed by the clustering algorithm (cophenetic distances) (Sokal and Rohlf, 1962), with the original Euclidean distances among objects. Of the evaluated clustering algorithms, average linkage gave the highest Pearson correlation (0.71) between the Euclidean distances among objects and the cophenetic distances among clusters, and was therefore considered appropriate. Average linkage minimizes the average of the pairwise distances between objects in different clusters (Sneath and Sokal, 1973). To assess the stability of the clustering, we calculated bootstrap probability values (BP values) for each cluster using the R package *pvclust* (Suzuki and Shimodaira, 2006) and performing 10,000 resamplings of the variables over all objects. The number of nodes, or branching points, in the resulting dendrogram along the path separating co-twins was then used as a measure of coclustering of co-twins (see Fig. 3A for an example). For each number of nodes separating co-twins, we compared the number of observations in the original clustering dendrogram with artificial situations where there is no clustering. In a dendrogram, the "root" of the tree (for example, in Fig. 2, the "top" of the dendrogram) is where all clusters ultimately merge, whereas each of the "leaves" at the "bottom" of the dendrogram corresponds to a single object, which is in our study a scaled average lipid profile of one individual. We created artificial negative control situations by 1000,000-fold Monte Carlo resampling of the object labels over the leaves of the observed clustering tree. For each of the individual permutations, the number of occasions where co-twins were separated by a given number of nodes in the clustering dendrogram was recorded. When the observed number of occasions where co-twins were separated by a given number of nodes in the dendrogram was above the 95% level of the distribution for that number of nodes as resulting from all permutation tests, the observed number of occasions for that number of separating nodes was considered statistically significant. Based upon this analysis, two subgroups of twins were identified, clustering either closely or not closely with their co-twin. For each case where co-twins did not cluster closely, we evaluated several participant characteristics and environmental factors that could provide an explanation for this.

## RESULTS

*Participants*

The total study cohort consisted of 54 participants from 23 families (30 males and 24 females), where 24 participants belonged to MZ male twin pairs (MZM) and 18 to MZ female twin pairs (MZF). One male–male and one female–female pair who were found to be DZ after additional genotyping (DZM and DZF; encoded as R_□ and F_○; see the legend to Fig. 2 for denotation of individual participants) were also included in the study. From seven families, a twin pair and a sibling of the same sex participated (three MZM, one DZM, two MZF, and one DZF). In one additional MZM family a female sibling (H_●) participated. The average age of the twins was 18.0 years (SD 0.2) and of the siblings 17.4 years (SD 4.3). Ac-
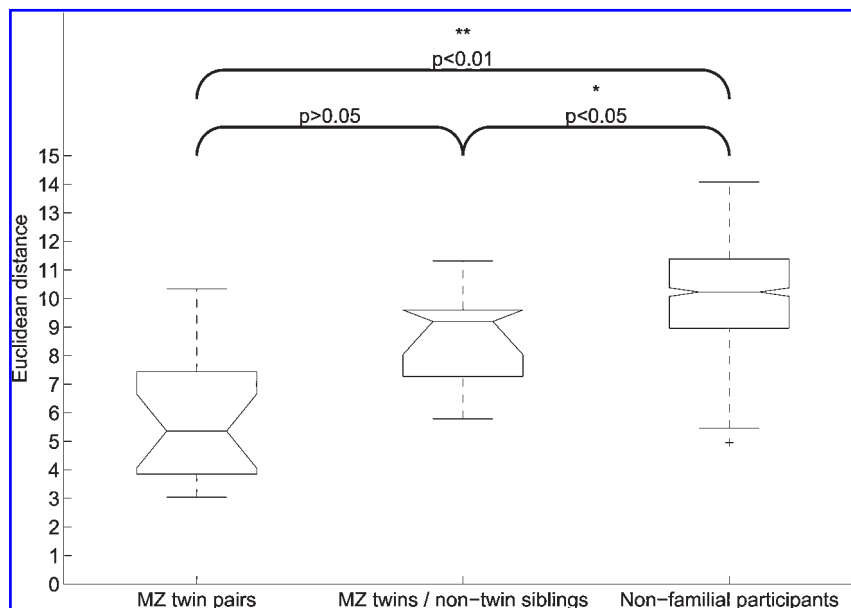
**FIG. 1.** Box-whisker plots showing distributions of Euclidean distances between MZ co-twins ($I/I$ = 38, left), between MZ twins and their same-sex nontwin siblings ($I/I$ = 10, middle), and among same-sex nonfamilial participants ($I/I$ = 637, right). Data from co-twins of twins in whose blood plasma samples we had noticed hemolysis (A_○_1 and X_□_2), as well as data from two DZ twin pairs (F_○ and R_□) were included in the computation of distances among nonfamilial subjects only. $p$-Values as resulting from a multiple comparison test of the group medians are displayed. *$p$ < 0.05; **$p$ < 0.01.

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

*Lipid profiling and data processing*

Blood plasma samples were analyzed with LC-MS, yielding profiles of 61 individual lipids per sample, which are listed in Figure 3C. The RSDs of the internal standard-corrected responses for the individual lipids in the quality control samples ranged from 5.2% to 25.5%. Notably, the RSDs of all LPCs, PCs, and SPMs were below 15%.

*Statistical analysis*

After averaging of the analytical duplicates and scaling of the data table, Euclidean distances between the 54 rows ("objects") were computed. The median within-pair distance for MZ twins was significantly smaller than the median distance among nonfamilial participants; the median distance between twins and their same-sex nontwin siblings was also significantly smaller than the median distance among same-sex nonfamilial participants (Fig. 1). Similar differences have been observed by Nanki and colleagues for gene expression, which, compared to the metabolome, is of course expected to correlate more strongly with genotype because it is less subject to environmental influences (Nanki et al., 1996). For the three investigated gene families, they found that the similarities in expression in peripheral blood lymphocytes were higher
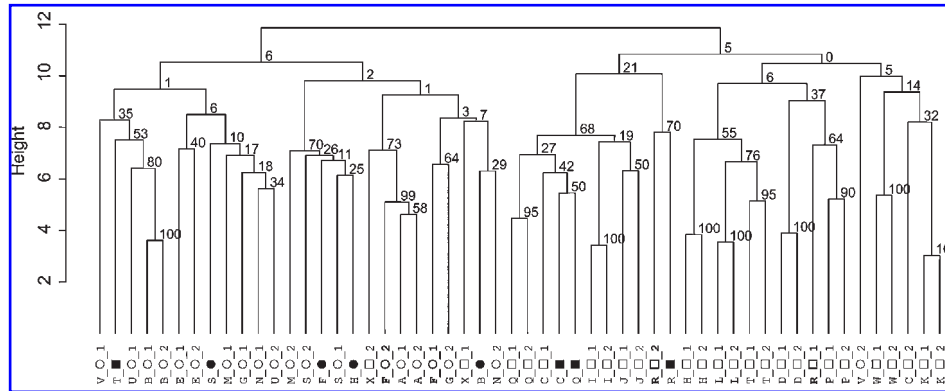
**FIG. 2.** Result of nonparametric bootstrap procedure. Numbers near the branching points in the dendrogram indicate BP values on basis of the data resampling method as explained. Denotation of participants: family, alphabetical letter (A–X); sex, squares for males and circles for females (e.g., [□] and [○] for a male and a female twin, respectively); 1,2, randomly allocated to individuals of a twin pair. Labels in bold type indicate DZ twin pairs. Nontwin siblings are indicated by filled squares (■) or filled circles (●) for males and females, respectively.

between MZ co-twins than among nonfamilial participants. The result of hierarchical clustering can be displayed as a tree, or dendrogram (Fig. 2) that denotes the relationships among clusters in a two-dimensional form. Female and male study participants are almost perfectly separated at the highest level. The dendrogram demonstrates considerable coclustering of MZ twin pairs. However, both DZ twin pairs do not cluster adjacently. Most nontwin siblings do not cluster closely with a sibling who is member of a twin pair. There appear to be rather few clusters that are either extremely tight or extremely loose. Figure 3A indicates with a color code Euclidean distances between all pairs of objects. The strong clustering of female (the upper left quadrant in Fig. 3A) and of male study participants (the lower right quadrant in Fig. 3A) is evident. In Figure 3C, the scaled data is shown for every participant as a separate vertical lane of the heatmap. The order of the objects along the horizontal axis is equal to that in Figure 3A. Again, panel C indicates that lipid profiles are different for males and females. The five lipid classes each coincide with a distinct pattern in the heatmap when viewing across all participants from top to bottom. Furthermore, this panel suggests that in general the TGs differentiate less than the other classes between samples from different families. LPCs and SPMs seem to differentiate most. Interestingly, there seem to be differences between families regarding the specific lipid compounds which are most similar among family members. The stability of the clustering of participants was assessed by a nonparametric bootstrap procedure. For a discussion of the result of this analysis we revert to Figure 2. In the context of hierarchical clustering, a bootstrap procedure can be used to investigate to which degree the dendrogram topology changes upon omitting or multiple occurrence of a number of variables for all objects. The stability of the clustering tends to be highest at the lowest level of clustering, that is, where the distance between clusters is relatively small. For the co-twins forming close clusters, BP values were in the range between 40 and 100, and in general clusters containing female co-twins had lower BP values than clusters of male co-twins. Therefore, especially for the female co-twins forming close clusters there may be subsets of variables that are especially important for the clustering. With this in mind, one way to improve the coclustering of twins may be to use a measure of object similarity, that is, COSA (Damian et al., 2007; Friedman and Meulman, 2004), that acknowledges that different subsets of objects may cluster on different subsets of variables. As a measure of coclustering of co-twins, for each twin pair we counted the number of branching points, or nodes, along the path separating both twins. The colors and heights of the lines that connect twins in Figure 3B indicate these numbers. For example, there were three twin pairs where the number of nodes between the co-twins was seven. In the dendrogram of Figure 3A an example is drawn of this characterization of the relative similarity of co-twins for a case where the distance between the co-twins is five nodes. For each possible number of nodes separating MZ or DZ twins, the observed frequency is displayed as a black dot in Figure 4. Characterizing the clustering of the nontwin siblings with their closest twin brother or sister in a similar
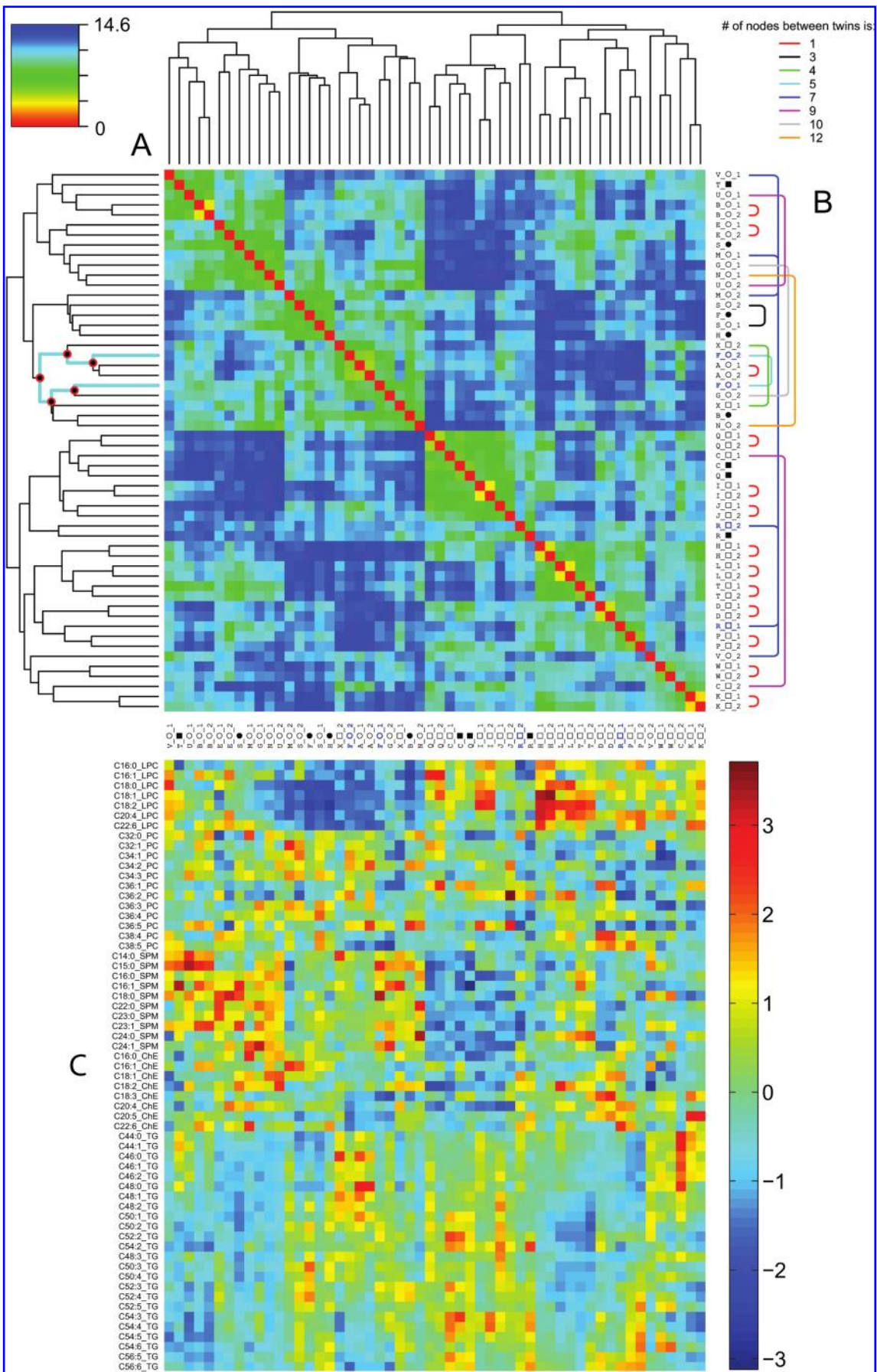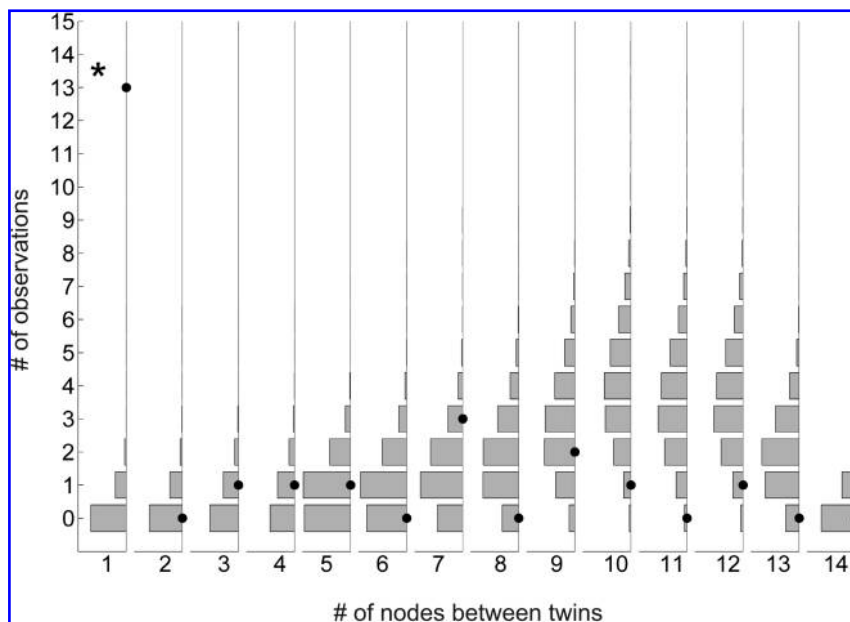
**FIG. 4.** Coclustering of twins compared with the results of permutation testing. Numbers of nodes separating co-twins increase from left to right. For each number of branching points, from bottom to top the number of twin pairs separated by that particular number of branching points after each permutation is displayed by gray bars; the number of observations in the original dendrogram (see Fig. 2 and Fig. 3A/B) is indicated by black dots. The asterisk (*) in the most left graph indicates that the observed number of 13 occasions where co-twins were separated by only one branching point is significantly different from what was observed in the permutation tests.

way, we found that five of these pairs of family members (i.e., B_●; F_■; H_●; S_●; and T_■ and their closest twin siblings) were separated by more than six nodes, and therefore did not cluster closely. We acknowledge that one difficulty with our approach is that the numbers of branching points along the path separating pairs of objects are not necessarily representative of the absolute magnitude of the dissimilarity between objects, in our case defined by Euclidean distance. For example, co-twins may be dissimilar in terms of Euclidean distance but still be separated by a limited number of nodes. This indicates that although they are dissimilar, they are still more similar to each other than to any other object in their neighborhood within the multidimensional space put up by the lipid profiles of all study participants. Thus characterizing the coclustering of twins in this way gives insight into the similarity of co-twins to each other, relative to the similarity of each individual twin with all other objects in the dataset. Subsequently, we tested whether coclustering of twins was indeed stronger than what would have been observed by chance, given the observed dendrogram topology. To this end, per possible number of nodes separating twins we created a reference distribution by permutation of the object labels over the leaves of the dendrogram. The significance of the observed numbers of nodes separating twins in the dendrogram was assessed by comparison with these reference distributions (Fig. 4). In this figure, from left to right, with each separate graph the number of branching points along the path in the dendrogram separating co-twins increases. Due to the given structure of the dendrogram, the maximum possible number of branching points along a path between two leaves was four-

◀

**FIG. 3.** Euclidean distances among objects and corresponding dendrogram (**A**); scaled data for each participant (**C**). In panel **B**, co-twins are connected by colored lines. In the dendrogram of panel **A** an example is drawn of our approach to characterize coclustering of twins. The keys to the colors in panels **A, B,** and **C** are given in the upper left, upper right, and lower right corners of the figure, respectively. In panel **C**, lipids are labeled by the number of carbon atoms as well as the number of double bonds (separated by a colon) in the fatty acid, followed by their class abbreviation (LPC, PC, . . . ).

TABLE 1. TENTATIVE EXPLANATIONS FOR THE SEPARATION OF CO-TWINS BY MORE THAN
ONE NODE IN THE DENDROGRAMS OF FIGURE 2 AND FIGURE 3A/B

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

teen, and therefore, the number of graphs in Figure 4 is also 14. For each number of branching points, from bottom to top the number of twin pairs separated by that particular number of branching points after each permutation is displayed by gray bars. In addition, for each possible number of branching points separating co-twins, the number of twin pairs separated by that number of nodes in the original dendrogram (see Fig. 2 and Fig. 3A/B) is indicated by black dots. For example, in Figure 3A/B, 13 twin pairs can be observed that are separated by only one node. Hence, in Figure 4 there is a black dot in the most left graph corresponding with one node separating twins, at the point corresponding with 13 twin pairs. In most permutations, no object labels of co-twins were separated by only one node, and therefore, the horizontal gray bar corresponding with zero observations in the same graph is tallest. As no twin pairs can be observed in Figure 3A/B where the number of nodes between twins is two, in the second graph from the left in Figure 4 there is a black dot corresponding with zero pairs, and so on. Using the results of the permutation tests, it was found that the observed number of 13 occasions where co-twins were only separated by one node was significantly different from what would have been observed by chance. For larger numbers of nodes separating co-twins, the observations with the object labels in original order fell within the distributions observed after the permutations. Therefore, we named "close" those co-twins who were separated by one node within the clustering tree and "distant" those co-twins who were separated by more than one node. The notion that there were two subgroups of either close or distant twins in the data, was supported by the observation that the distribution of the within-pair Euclidean distances partly overlapped with the distribution of distances among nonfamilial study participants, as was shown in Figure 1. For each "distant" twin pair, we have attempted to provide an explanation for the observed separation of the co-twins by more than one node
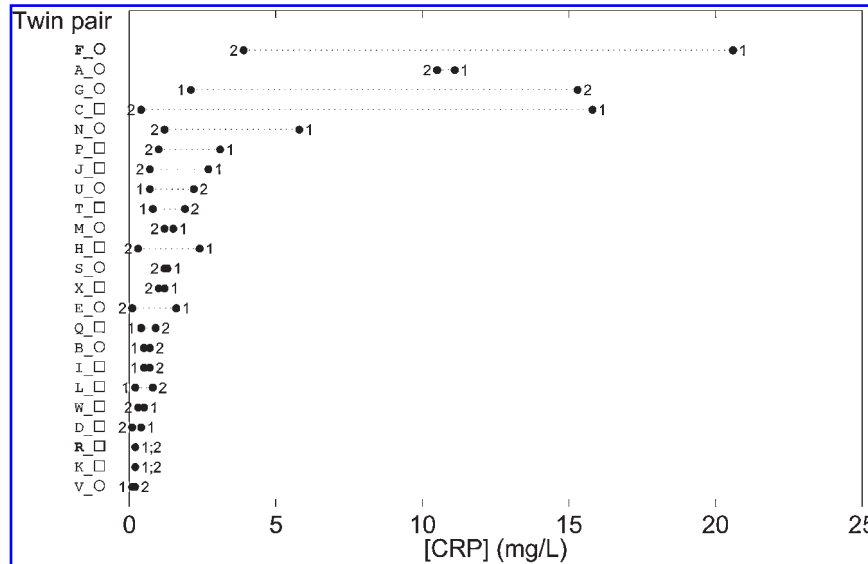
**FIG. 5.** CRP levels in blood samples in twins. From bottom to top, the average CRP level of twin pairs increases. The numbers "1" and "2" near the observations denote the "first" and "second" twin of each pair, respectively. For an explanation of this labeling, see the legend of Figure 2.

in the dendrogram (Table 1). These explanations were based upon the available information on participant characteristics and environmental factors. Moreover, in a number of cases dissimilarity of lipid profiles correlated with within-pair differences in the levels of the inflammatory marker CRP (Fig. 5). In particular, female sex and recent illness correlated with dissimilarity of lipid profiles between MZ co-twins. In turn, in a number of cases, recent illness as self-reported by the study participants correlated with an increased level of CRP. However, we could not establish the influence of female sex and recent illness independently, because a relatively large number of female study participants had self-reportedly been ill. Moreover, a number of female "distant" twin pairs did not have synchronous menstrual cycles. Dizygosity correlated strongly with dissimilarity of lipid profiles as well, as both DZ twin pairs included in the study were separated by more than one node in the dendrogram. Moreover, five out of a total of eight nontwin siblings included in the study—of whom all except H_● were of the same sex as their siblings belonging to a pair of twins—did not cluster closely with a sibling belonging to a pair of twins. This observation suggests that the dissimilarity of both DZ twin pairs was caused by differences in genetic background rather than by differences in environmental factors. That is, if shared environment would have been more important for the similarity of lipid profiles, the similarity of nontwin siblings with their twin siblings would have approached the similarity of MZ twin pairs. Although the relative within-pair dissimilarity of lipid profiles correlated with differences in genetic background or environmental exposure, some twin pairs had relatively similar lipid profiles despite the presence of such differences. For example, twins J_□_1 and D_□_1 had had a cold less than 1 week prior to blood sampling whereas their co-twins had not. Still, both discordant pairs were not found to be distant in the clustering. Also, none of the female "close" twin pairs did have completely synchronous menstrual cycles.

## DISCUSSION

In this study we have shown that upon hierarchical clustering of lipid profiles from healthy MZ twins, a significant number of co-twins forms close clusters. This is a strong indication that similarities in genetic background and/or environmental history among individuals indeed manifest as similarities in lipid profiles. Where the genetic resemblance of family members is expected to be lower than in MZ twin pairs on

basis of Mendelian inheritance, that is, between DZ twins and between twins and their nontwin siblings, we observed lower similarity of lipid profiles. Moreover, in a number of cases where MZ co-twins did not cluster closely, we have identified recent experiences that might have decreased the within-pair similarity, suggesting an important role of environmental influences in these pairs. Indeed, the similarities among non-familial participants, who are expected to share less genetic background and environmental exposure than family members, were low on average. To our knowledge, this is one of the first reports on unsupervised data analysis of metabolite profiles among healthy twins. Until now, in most publications gene expression data were studied. For example, Tan et al. (2005) applied their correspondence analysis to project gene expression, measured using whole blood mRNA, separately for each of the 12 elderly female MZ and DZ twins included in their study. They observed that in two MZ and two DZ twin pairs the within-pair correlation of gene expression was higher compared to the correlation between twins from different pairs. Moreover, in these four pairs the within-pair correlation in expression in the MZ pairs was higher than that in the DZ pairs. Omori-Inoue et al. (2007) performed hierarchical clustering based on correlations between gene expression profiles in umbilical cords from five twin pairs, and found that in four—probably MZ—pairs the co-twins clustered adjacently, whereas the non-adjacently clustering co-twins (the fifth pair) might have been DZ. In the result of hierarchical clustering on basis of 102 genes differentially expressed in skin fibroblasts from study participants with systemic sclerosis compared to controls, two MZ twin pairs were observed of which the co-twins—who were discordant for the disease—clustered adjacently (Zhou et al., 2005). Matigian et al. (2007) used Pearson correlation as the similarity measure for hierarchical clustering of gene expression profiles in lymphoblastoid cell lines from three MZ twin pairs that were discordant for bipolar disorder, and found that the co-twins of all pairs clustered adjacently. Such high within-pair similarity in MZ pairs was not observed by Teuffel et al. (2004), who subjected the bone marrow gene expression profiles of 33 children with acute lymphoblastic leukemia, including one pair of MZ twins concordant for the disease, to hierarchical clustering using Euclidean distances and applying the average linkage clustering algorithm. The authors noticed that the co-twins did not cluster adjacently and ascribed this effect to disease-related changes in gene expression. Two recent articles employ supervised methods to analyze metabolomics data from twins. In one publication a link was established between schizophrenia and alterations in blood plasma lipid levels as assessed by ^{1}H NMR spectroscopy (Tsang et al., 2006). Such changes were observed in both male and female affected twins of pairs discordant for schizophrenia when compared to age-matched control twin pairs. However, in females, the differences between the affected twins and their control twins were more pronounced than in males; as opposed to in males, in females the authors also observed a significant difference between the unaffected twins of discordant pairs and control twins. The larger effects in females were attributed to greater genetic predisposition to the disease-related changes in discordant female pairs than in male pairs. A recent study by Pietilainen and colleagues (2007) found within-pair differences in lipid profiles, as assessed in blood serum using LC-MS, in MZ pairs discordant for obesity. Interestingly, the authors report that compared with five normal-weight concordant as well as with five pairs concordant for overweight, the discordant pairs did not have larger intrapair differences in total cholesterol, high-density lipoprotein, low-density lipoprotein or TGs. Our results suggest that an unsupervised data analysis approach (Hastie et al., 2001) can yield information that can not be derived from other, pseudosupervised analyses. In explorative studies like this one, any constraints in supervised analysis may preclude novel findings. Using hierarchical cluster analysis, we were able to link within-pair dissimilarity of lipid profiles to within twin pair-specific factors. Studies estimating the relative influence of genetic variation on the within-twin pair variation in lipid levels, have been reviewed by Iselius (1979), and by Snieder et al. (1999). To our knowledge, with respect to the lipid classes evaluated in this study, only heritability estimates for the TGs have been described previously. A study based upon a population sample having a mean age of 16.7 years, which is close to the mean age of our study cohort, found that genetic factors accounted for 60% of the variation in total TG levels among individuals (Boomsma et al., 1996). In general, the relative influence of genetic variation on the phenotypic variation in lipid and apolipoprotein levels has been found to be high. Such high estimates are consistent with our findings, of successful clustering of individual twin pairs based upon unsupervised analysis of phenotypic data. In addition to genetic background and environmental factors, experimental factors such as sample handling and storage, as well as the used analytical methods may introduce further similarities and differences between

lipid profiles from different individuals. In our study, lipid profiles were assessed in blood plasma samples from fasting individuals, because during fasting lipid profiles are thought to be relatively stable (Cohn et al., 1988), and therefore expected to be more similar among individuals sharing genetic makeup and/or environmental exposure. Although we can not completely exclude the possibility that the similarities among samples from MZ twin pairs are partly due to other shared factors induced by the study setup, the larger dissimilarities between twins and their nontwin siblings argues against a strong influence of such factors. Samples from members of the same family (i.e., twins and additional siblings) were collected on the same day. If the workup of samples from a given family would have introduced similarities among the samples from that family relative to samples from other families, a larger resemblance of twin-sibling pairs would have been observed. With respect to sample handling and storage, we suspect that hemolysis of blood samples may augment differences in lipid profiles. In one out of two cases where noticeable hemolysis of the blood sample had occurred, the corresponding twin pair was found to be separated by more than one node in the dendrogram. We found that healthy individuals who share genetic background and/or environmental exposure, have blood plasma lipid profiles that are more similar than profiles of persons who do not share these influences. When extending this observation in twins to a general healthy population, this probably implies that the lipid profile corresponding with a healthy state is characteristic for each individual due to the individual-specific genetic background and environmental exposure (German et al., 2003). We therefore suspect that changes in the lipid profile might denote deviations from the healthy phenotype, and therefore could be used, for example, to diagnose the onset of disease. The correlation of an increased blood concentration of the inflammatory marker CRP with dissimilarity of lipid profiles in a number of MZ twins in the current study supports this hypothesis. Actually, it can be assumed that for each individual there is a lipid profile describing the "healthy phenotype", and in the context of personalized medicine the aim could be to maintain this, or to take measures to restore it. In conclusion, in our study, healthy MZ twins have relatively similar blood plasma lipid profiles. Between individuals with less shared genetic backgrounds and environmental exposure, we indeed observed smaller similarity. Discordance of MZ twins for recent disease, that can be regarded a particularly relevant difference in environmental exposure, correlated well with within-pair dissimilarity of lipid profiles. Therefore, lipid profiling might prove useful in monitoring personal health.

## ACKNOWLEDGMENTS

## REFERENCES

ACTON, S., RIGOTTI, A., LANDSCHULZ, K.T., XU, S., HOBBS, H.H., and KRIEGER, M. (1996). Identification of scavenger receptor SR-BI as a high density lipoprotein receptor. Science **271,** 518–520.

BARNES, R.J., DHANOA, M.S., and LISTER, S.J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl Spectrosc **43,** 772–777.

BAUR, L.A., O'CONNOR, J., PAN, D.A., WU, B.J., O'CONNOR, M.J., and STORLIEN, L.H. (2000). Relationships between the fatty acid composition of muscle and erythrocyte membrane phospholipid in young children and the effect of type of infant feeding. Lipids **35,** 77–82.

BOOMSMA, D.I., KEMPEN, H.J., GEVERS LEUVEN, J.A., HAVEKES, L., DE KNIJFF, P., and FRANTS, R.R. (1996). Genetic analysis of sex and generation differences in plasma lipid, lipoprotein, and apolipoprotein levels in adolescent twins and their parents. Genet Epidemiol **13,** 49–60.

BOOMSMA, D.I., DE GEUS, E.J., VINK, J.M., STUBBE, J.H., DISTEL, M.A., HOTTENGA, J.J., et al. (2006). Netherlands Twin Register: from twins to twin families. Twin Res Hum Genet **9,** 849–857.

COHN, J.S., MCNAMARA, J.R., COHN, S.D., ORDOVAS, J.M. and SCHAEFER, E.J. (1988). Postprandial plasma lipoprotein changes in human subjects of different ages. J Lipid Res **29,** 469–479.

COLEMAN, R. (1987). Biochemistry of bile secretion. Biochem J **244,** 249–261.

DAMIAN, D., ORESIC, M., VERHEIJ, E.R., MEULMAN, J.J., FRIEDMAN, J.H., ADOURIAN, A., et al. (2007). Applications of a new subspace clustering algorithm (COSA) in medical systems biology. Metabolomics **3,** 69–77.

DODDS, P.F. (1991). Incorporation of xenobiotic carboxylic acids into lipids. Life Sci **49,** 629–649.

ELSTAD, M.R., STAFFORINI, D.M., MCINTYRE, T.M., PRESCOTT, S.M., and ZIMMERMAN, G.A. (1989). Platelet-activating factor acetylhydrolase increases during macrophage differentiation. A novel mechanism that regulates accumulation of platelet-activating factor. J Biol Chem **264,** 8467–8470.

FISCHER, K., BOT, A.N., ZWAAN, B.J., and BRAKEFIELD, P.M. (2004). Genetic and environmental sources of egg size variation in the butterfly Bicyclus anynana. Heredity **92,** 163–169.

FRIEDMAN, J.H., and MEULMAN, J.J. (2004). Clustering objects on subsets of attributes. J R Stat Soc Series B **66,** 815–849.

GERMAN, J.B., ROBERTS, M.A., and WATKINS, S.M. (2003). Personal metabolomics as a next generation nutritional assessment. J Nutr **133,** 4260–4266.

GLOMSET, J.A. (1968). The plasma lecithins:cholesterol acyltransferase reaction. J Lipid Res **9,** 155–167.

GOODACRE, R. (2007). Metabolomics of a superorganism. J Nutr **137,** 259S–266S.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J.H. (2001). *The Elements of Statistical Learning; Data Mining, Inference, and Prediction* (Springer, New York).

HOCHBERG, Y., and TAMHANE, A.C. (1987). *Multiple Comparison Procedures* (John Wiley & Sons, New York).

HOEKSTRA, R.A., BARTELS, M., and BOOMSMA, D.I. (2007). Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. Learn Individ Differ **17,** 97–114.

ISELIUS, L. (1979). Analysis of family resemblance for lipids and lipoproteins. Clin Genet **15,** 300–306.

KERN, H., VOLK, T., KNAUER-SCHIEFER, S., MIETH, T., RUSTOW, B., KOX, W.J., et al. (1998). Stimulation of monocytes and platelets by short-chain phosphatidylcholines with and without terminal carboxyl group. Biochim Biophys Acta **1394,** 33–42.

KOEK, M.M., MUILWIJK, B., VAN DER WERF, M.J., and HANKEMEIER, T. (2006). Microbial metabolomics with gas chromatography/mass spectrometry. Anal Chem **78,** 1272–1281.

KRUSKAL, W.H., and WALLIS, W.A. (1952). Use of ranks in one-criterion variance analysis. J Am Stat Assoc **47,** 583–621.

MATIGIAN, N., WINDUS, L., SMITH, H., FILIPPICH, C., PANTELIS, C., MCGRATH, J., et al. (2007). Expression profiling in monozygotic twins discordant for bipolar disorder reveals dysregulation of the WNT signalling pathway. Mol Psychiatry **12,** 815–825.

MCKEONE, B.J., PATSCH, J.R., and POWNALL, H.J. (1993). Plasma triglycerides determine low density lipoprotein composition, physical properties, and cell-specific binding in cultured cells. J Clin Invest **91,** 1926–1933.

MURPHY, D.J. (2001). The biogenesis and functions of lipid bodies in animals, plants and microorganisms. Prog Lipid Res **40,** 325–438.

NANKI, T., KOHSAKA, H., MIZUSHIMA, N., OLLIER, W.E., CARSON, D.A., and MIYASAKA, N. (1996). Genetic control of T cell receptor BJ gene expression in peripheral lymphocytes of normal and rheumatoid arthritis monozygotic twins. J Clin Invest **98,** 1594–1601.

NEALE, M.C., and CARDON, L.R. (1992). *Methodology for Genetic Studies of Twins and Families* (Kluwer Academic Publishers, Dordrecht).

OMORI-INOUE, M., FUKATA, H., KOMIYAMA, M., TODAKA, E., OSADA, H., ABURATANI, H., et al. (2007). The contamination levels of organochlorines and the pattern of gene expressions in human umbilical cords from intra-pairs of twins at delivery. Reprod Toxicol **23,** 283–289.

PIETILAINEN, K.H., SYSI-AHO, M., RISSANEN, A., SEPPANEN-LAAKSO, T., YKI-JARVINEN, H., KAPRIO, J., et al. (2007). Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects—A monozygotic twin study. PLoS ONE **2,** e218.

QUINN, M.T., PARTHASARATHY, S., and STEINBERG, D. (1988). Lysophosphatidylcholine: a chemotactic factor for human monocytes and its potential role in atherogenesis. Proc Natl Acad Sci USA **85,** 2805–2809.

R DEVELOPMENT CORE TEAM. (2005). R: *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

SIGAL, Y.J., MCDERMOTT, M.I., and MORRIS, A.J. (2005). Integral membrane lipid phosphatases/phosphotransferases: common structure and diverse functions. Biochem J **387,** 281–293.

SNEATH, P.H.A., and SOKAL, R.R. (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification* (W.H. Freeman & Co., New York).

SNIEDER, H., VAN DOORNEN, L.J., and BOOMSMA, D.I. (1999). Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies. Arterioscler Thromb Vasc Biol **19,** 2826–2834.

SOKAL, R.R. (1961). Distance as a measure of taxonomic similarity. Syst Zool **10,** 70–79.

SOKAL, R.R. and ROHLF, F.J. (1962). The comparison of dendrograms by objective methods. Taxon **11,** 33–40.

SUZUKI, R., and SHIMODAIRA, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics **22,** 1540–1542.

SWIFT, R.W. (1932). The effects of low environmental temperature upon metabolism: II. The influence of shivering, subcutaneous fat, and skin temperature on heat production. J Nutr **5,** 227–249.

TAN, Q., CHRISTENSEN, K., CHRISTIANSEN, L., FREDERIKSEN, H., BATHUM, L., DAHLGAARD, J., et al. (2005). Genetic dissection of gene expression observed in whole blood samples of elderly Danish twins. Hum Genet **117,** 267–274.

TEUFFEL, O., BETTS, D.R., DETTLING, M., SCHAUB, R., SCHAFER, B.W., and NIGGLI, F.K. (2004). Prenatal origin of separate evolution of leukemia in identical twins. Leukemia **18,** 1624–1629.

TEW, D.G., SOUTHAN, C., RICE, S.Q., LAWRENCE, M.P., LI, H., BOYD, H.F., et al. (1996). Purification, properties, sequencing, and cloning of a lipoprotein-associated, serine-dependent phospholipase involved in the oxidative modification of low-density lipoproteins. Arterioscler Thromb Vasc Biol **16,** 591–599.

TSANG, T.M., HUANG, J.T., HOLMES, E., and BAHN, S. (2006). Metabolic profiling of plasma from discordant schizophrenia twins: correlation between lipid signals and global functioning in female schizophrenia patients. J Proteome Res **5,** 756–760.

VANDEGINSTE, B.G.M., MASSART, D.L., BUYDENS, L.M.C., DE JONG, S., LEWI, P.J., and SMEYERS-VERBEKE, J. (1998). *Handbook of Chemometrics and Qualimetrics: Part B* (Elsevier, Amsterdam).

VAN DER GREEF, J., DAVIDOV, E., VERHEIJ, E.R., VOGELS, J.T., VAN DER HEIJDEN, R., ADOURIAN, A., et al. (2004). The role of metabolomics in systems biology. In: *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. G. G. Harrigan and R. Goodacre, eds. (Kluwer Academic Publishers, Boston), pp. 171–198.

VELDHUIZEN, R., NAG, K., ORGEIG, S., and POSSMAYER, F. (1998). The role of lipids in pulmonary surfactant. Biochim Biophys Acta **1408,** 90–108.

YOUNG, G., and HOUSEHOLDER, A.S. (1938). Discussion of a set of points in terms of their mutual distances. Psychometrika **3,** 19–22.

ZACHOWSKI, A. (1993). Phospholipids in animal eukaryotic membranes: transverse asymmetry and movement. Biochem J **294**(Pt 1), 1–14.

ZHOU, X., TAN, F.K., XIONG, M., ARNETT, F.C., and FEGHALI-BOSTWICK, C.A. (2005). Monozygotic twins clinically discordant for scleroderma show concordance for fibroblast gene expression profiles. Arthritis Rheum **52,** 3305–3314.

Address reprint requests to:
*Thomas Hankemeier*
*Leiden University, LACDR*
*P.O. Box 9502*
*NL-2300 RA Leiden*
*The Netherlands*

*E-mail:* hankemeier@chem.leidenuniv.nl