European Journal of Human Genetics (2013) 21, 95–101

www.nature.com/ejhg

ARTICLE

Hierarchical clustering analysis of blood plasma lipidomics profiles from mono- and dizygotic twin families

Harmen HM Draisma*[,1,5], Theo H Reijmers[1], Jacqueline J Meulman[2], Jan van der Greef[1], Thomas Hankemeier[1,3,6] and Dorret I Boomsma[4,6]

Twin and family studies are typically used to elucidate the relative contribution of genetic and environmental variation to phenotypic variation. Here, we apply a quantitative genetic method based on hierarchical clustering, to blood plasma lipidomics data obtained in a healthy cohort consisting of 37 monozygotic and 28 dizygotic twin pairs, and 52 of their biological nontwin siblings. Such data are informative of the concentrations of a wide range of lipids in the studied blood samples. An important advantage of hierarchical clustering is that it can be applied to a high-dimensional 'omics' type data, whereas the use of many other quantitative genetic methods for analysis of such data is hampered by the large number of correlated variables. For this study we combined two lipidomics data sets, originating from two different measurement blocks, which we corrected for block effects by 'quantile equating'. In the analysis of the combined data, average similarities of lipidomics profiles were highest between monozygotic (MZ) cotwins, and became progressively lower between dizygotic (DZ) cotwins, among sex-matched nontwin siblings and among sex-matched unrelated participants, respectively. Our results suggest that (1) shared genetic background, shared environment, and similar age contribute to similarities in blood plasma lipidomics profiles among individuals; and (2) that the power of quantitative genetic analyses is enhanced by quantile equating and combination of data sets obtained in different measurement blocks.
European Journal of Human Genetics (2013) **21,** 95–101; doi:10.1038/ejhg.2012.110; published online 20 June 2012

Keywords: hierarchical clustering analysis; twin study; metabolomics

INTRODUCTION

Genetic variation and variation in environmental influences among individuals contribute to individual differences in measurable characteristics, that is, 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[1,2] of traits, often genetically informative samples such as (twin) families are used. Compared with heritability analyses in nuclear families, studies on the basis of twin families have enhanced power to detect genetic influences on phenotypic variation.[3,4] One cause for this is that the members of twin pairs living together are particularly well-matched for age, prenatal factors and environmental variation. A second cause is the difference in shared genetic variation between MZ and DZ twin pairs. MZ twins share all genetic variation. DZ twins, on the other hand, share only approximately half of the variation at the DNA sequence level, and the same degree of genetic variation is shared among biological nontwin siblings.[5] Because of this large difference in shared genetic variation between MZ and DZ twin pairs, and the strong matching for environmental variation within both types of twin pairs, comparison of the phenotypic correlations between MZ and DZ twin pairs enables estimation of heritability with relatively high power. This type of quantitative genetic analysis is often carried out using structural equation modeling (SEM).[5,6]

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[7–9] 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 that incorporate endophenotypes are potentially informative of the biological pathways leading to the observed phenotypic variation among individuals. Among the endophenotypes, metabolite levels are particularly interesting because metabolites are relatively close to the phenotype and therefore potentially directly relevant for phenotypic variation. Metabolomics aims at the holistic analysis of the substrates, intermediates and end products of cellular metabolism present in biological samples.[10] By using advanced analysis methods such as

[1]Division Analytical Biosciences, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands; [2]Leiden University, Mathematical Institute, Leiden, The Netherlands; [3]Netherlands Metabolomics Centre, Leiden, The Netherlands; [4]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands
*Correspondence: Dr HHM Draisma, Department of Biological Psychology and the EMGO+ Institute for Health and Care Research, Faculty of Psychology and Education, VU University Amsterdam, Van der Boechorststraat 1, room 2B-09, NL-1081 BT Amsterdam, The Netherlands. Tel: +31 20 5986871; Fax: +31 20 5988832; E-mail: h.h.m.draisma@vu.nl
[5]Present address: Department of Biological Psychology and the EMGO+ Institute for Health and Care Research, Faculty of Psychology and Education, VU University Amsterdam, Van der Boechorststraat 1, room 2B-09, NL-1081 BT Amsterdam, The Netherlands
[6]These authors contributed equally to this work.
Received 1 February 2012; revised 25 April 2012; accepted 2 May 2012; published online 20 June 2012

proton nuclear magnetic resonance spectroscopy or liquid chromatography coupled to mass spectrometry (LC–MS), it is possible to detect the individual metabolites that belong to a particular metabolite class rather than the conventional biochemical summary measures such as 'total triglycerides'.[11,12] Because of their relatively unbiased, comprehensive nature, metabolomics studies allow for a more direct linking of findings to 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 (ie, 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.[13] Information regarding group membership is not used during the clustering process; rather, objects that have similar scores on the 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 two-dimensional structure. When performing HCA of multivariate data obtained in different families, because of the genetic and environmental variation shared by family members it is expected that members of the same family will cluster. Members of different families, on the other hand, are expected to be in different clusters. MZ twins of the same pair are expected to cluster very strongly because they are genetically identical.

A useful property of HCA in general is that the sample covariance matrix does not have to be invertible. Therefore, in contrast to for example, maximum likelihood-based SEM, HCA can straightforwardly be applied for quantitative genetic analysis based on typical 'omics' data that consist of large numbers of correlated variables.

In the context of (twin) family studies, another advantage of HCA is that it acknowledges the pleiotropic effects of genes influencing the variation of different traits belonging to the same biological pathway. For example, if the blood plasma levels of different metabolites covary because these metabolites are subject to conversion by the same metabolic enzyme, in HCA the effects in all metabolites will be pooled. This then contributes to the statistical power of the analysis. 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 an earlier publication[14] we demonstrated that in HCA of blood plasma lipidomics data obtained in a sample consisting mainly of MZ twin pairs, male and female study participants were separated at the highest level in the clustering dendrogram. This suggested that variation in lipidomics profiles is relatively small among individuals of the same gender.

In the current paper, we report the results of HCA of blood plasma lipidomics data from an unselected general population cohort of 37 MZ twin pairs, 28 DZ twin pairs, and in total 52 of their biological nontwin siblings. Lipidomics, that is, the analysis of lipids with metabolomics techniques, is an important branch of metabolomics because lipids are involved in a plethora of (patho) physiological processes.[15] For the current study, we combined the data that provided the basis for our previous article[14] with additional data mainly from DZ twin pairs and from biological nontwin siblings. The initial data and the additional data were acquired with the same LC–MS method but in different measurement 'blocks', that is, in different sets of measurements consisting of one or more measurement batches. Therefore, we applied the method of 'quantile equating' to make the data from both measurement blocks combinable.[16] The inclusion in the current study of more DZ twin pairs and more nontwin siblings allowed for observation of the effect of age on similarity of lipidomics profiles, in addition to observation of the influences of genes and environment. Furthermore, the current paper demonstrates that the application of quantile equating to make combinable metabolomics data sets indeed causes biological effects to be visible in the combined data set, rather than nonbiological differences between the data from different measurement blocks.

SUBJECTS AND METHODS

Participants

Twins and biological nontwin siblings were recruited from the Netherlands Twin Register.[17] Characterization of participants, collection of fasting blood and urine samples, and sample preparation were performed as described previously.[14,16,18] Participants completed a number of surveys; 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 together. Female participants reported the day of their menstrual cycle at the time of blood sampling. Zygosity was determined for all twin pairs by DNA genotyping.

Measures

Lipidomics profiling and measurement of C-reactive protein (CRP) concentration in blood plasma samples were performed as described previously.[14,16] In brief, lipidomics profiling was performed using an LC–MS method targeted at the analysis of lipids that become positively charged during ionization in the mass spectrometer. These measurements were carried out in two consecutive 'blocks', denoted as B1 and B2, respectively. The measurements of B2 were performed almost one year after those of B1; samples from members of the same family were always measured in the same block. In B1, two replicate measurements were performed per study sample, whereas in B2 each study sample was measured only once.

The nonbiological systematic differences between the normalized data from the two measurement blocks were removed by 'quantile equating'.[16] In quantile equating, the distributions of the observations for the same metabolite in different data sets are averaged, assuming that all observations originate from the same population. This corrects the data for practically unavoidable slight technical differences between measurement blocks. For B1, we averaged the data from the two replicate measurements for each sample before equating.

Hierarchical clustering analysis

Cluster 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. 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.[14] Subsequently the lipidomics profiles were normalized among individuals (rows) by standard normal variate (SNV) normalization.[19] 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).[20] Euclidean distance matrices were subjected to HCA using the average linkage clustering algorithm, which among the evaluated clustering algorithms (average, median, single, complete and centroid linkage; Ward's and McQuitty's methods) gave the highest Pearson correlation between the original distances and the cophenetic distances among all study participants.[21] Heatmaps and associated hierarchical clustering dendrograms were generated using the 'heatmap.2' function in the 'gplots' package in the statistical computing environment R.[22] The remaining analyses were performed using the combined B1–B2 data set after quantile equating only. The distributions of the Euclidean distances between MZ cotwins, between DZ cotwins, among sex-matched nontwin siblings, and among sex-matched

unrelated participants were visualized by box plots. To assess whether there were statistically significant differences in median Euclidean distance among these groups, we performed a multiple comparison procedure based on the result of a nonparametric analysis of the variance within these groups of study participants versus the variance of the group medians and adopting Tukey's honestly significant difference criterion.[23] A multiple comparison procedure is designed to be conservative when testing for significant differences for more than one pair of groups.[24] The stability of the hierarchical clustering based on the distances among objects was assessed by a bootstrap analysis (10 000 resamplings) using the 'pvclust' package[25] 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 characterized by counting for each pair of same-sex relatives the number of separating nodes or branching points in the dendrogram[26,27] (see Supplementary Figure S1 for an example of this 'node analysis'). This number of separating nodes is representative of the relative proximity of individuals according to their lipid profiles, and hence acknowledges that the same Euclidean distance may have a different interpretation in sparsely and densely populated areas of the total multivariate space put up by the lipid profiles of all study participants.[14]

For each possible number of nodes separating MZ or DZ cotwins or same-sex nontwin siblings in the dendrogram, we compared the observed number of cotwin or sibling pairs separated by that number of nodes with the number of observations that was expected on the 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 SDs of these average *P*-values. For these comparisons, we used a critical value of 5% to denote statistical significance.

Association of node distances with Euclidean distance and with important covariates

Spearman's correlation coefficient was used to assess the correlation between the numbers of nodes in the dendrogram connecting pairs of individuals, with the original Euclidean distances among all study participants.

We performed linear regression analysis including several important covariates to test the association of these variables with the numbers of nodes separating sex-matched family members. The averaged log-transformed CRP concentration for each pair of relatives was included as a covariate in this regression as well, because in separate analysis of the B1 data this measure appeared to correlate with increased dissimilarity of lipid profiles between MZ cotwins.[14] Hence, the total set of included covariates comprised of: sex; pair average CRP level; measurement block; and monozygosity.

RESULTS

Participants

The participants originated from in total 65 families; 79 participants were male and 103 were female (Table 1). In one monozygotic female family and one DZ male family, a twin pair and two nontwin siblings (in both the 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.

Table 1 Basic description of participants

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

Abbreviations: DZF, dizygotic female; DZM, dizygotic male; MZF, monozygotic female; MZM, monozygotic male.

Measures

The combined data set, based on the measurements obtained in the two measurement blocks, comprised data on 59 lipids from five classes detected in the sample from each participant: lysophosphatidylcholines (LPCs; six lipids), phosphatidylcholines (twelve lipids), cholesterol esters (ChEs; ten lipids), sphingomyelins (eight lipids), and triglycerides (TGs; 23 lipids). One LPC and one TG displaying high measurement variance in B2 were excluded from the combined data set, as well as six additional LPCs, one TG and one ChE that had been reported only in the samples measured in B2.

Hierarchical clustering analysis

The results of HCA are displayed as dendrograms, together with the ordered heatmap that represents the Euclidean distance for each pair of objects (Figure 1). The Pearson correlations between the original Euclidean distances and the cophenetic distances based on HCA of the combined B1–B2 data sets were equal to 0.75 and 0.60 before and after equating, respectively (see Supplementary Table S1 for these correlations for the other evaluated clustering methods). Before quantile equating, the subjects in the combined (concatenated with the variables as the shared mode) B1–B2 data set clustered very strongly according to measurement block (Figure 1a). However, after equating, subjects measured in the two respective blocks were dispersed among each other (Figure 1b). The stability of the clustering of participants in 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 (Supplementary Figure S2). In concordance with our previous results using the separate B1 data before equating, in the combined equated B1–B2 data sets the average Euclidean distance appeared to increase when considering MZ cotwins, nontwin siblings, and unrelated participants, respectively (Figure 2). In addition, the distances between DZ cotwins appeared to assume a middle ground between those separating MZ cotwins and those separating nontwin siblings. Indeed, the differences in median Euclidean distance between several subgroups of participants were statistically significant according to a multiple comparison procedure (Table 2).

Clustering of MZ cotwins, of DZ cotwins, and of nontwin siblings in the combined equated B1–B2 data set were characterized using 'node analysis'. Spearman's correlation between the Euclidean distances for all pairs of study participants and their distances according to node analysis was equal to 0.44, $P<0.001$. 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 cotwins or nontwin siblings, with a reference distribution as provided by permutation testing. The results of these comparisons are visualized and summarized in Figure 3 and in Supplementary Table S2, respectively. For the MZ twin pairs only the number of occasions (in the current study fifteen) where cotwins 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 3a and Supplementary Table S2A). 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 3b and Supplementary Table S2B). 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

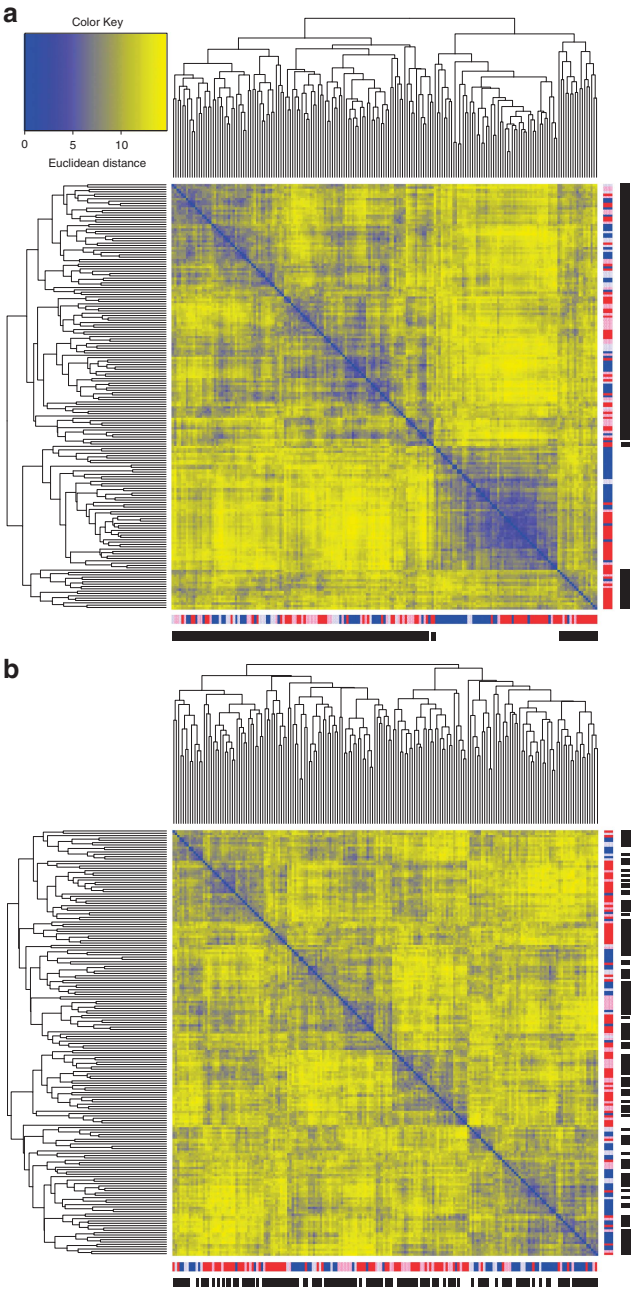

Figure 1 Heatmaps of Euclidean distances between objects, and associated hierarchical clustering dendrograms for combined B1–B2 data set before and after quantile equating. B1 and B2 data sets were combined by concatenation with the variables (lipids) as the shared mode. (**a**) Before quantile equating; (**b**) After 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). DZ female and DZ 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).

by three (two pairs), five (three pairs), eight (six pairs) or nineteen nodes (seven pairs) (Figure 3c and Supplementary Table S2C). Indeed, when including all pairs of sex-matched related participants in linear regression analysis, monozygotic cotwins were separated by significantly smaller numbers of nodes in the dendrogram compared with other types of pairs, $t(121) = -3.201$, $P = 0.002$. The effects of sex, measurement block and mean CRP level on the number of nodes separating pair members were not significant in this regression.

In Supplementary Tables S3 and S4, descriptions are given for MZ cotwins separated by only one and by more than one node in the combined equated B1–B2 data sets, respectively. Within the group of MZ twin pairs, female sex was significantly associated with an increased number of nodes separating cotwins when adjusting for measurement block and for CRP concentration averaged over both cotwins: $t(33) = 2.498$, $P = 0.02$. The regression coefficients for pair average CRP level and for measurement block were not significantly different from zero.

DISCUSSION

Influences of genes, environment and age

We report the results of application of HCA to multivariate data from a genetically informative sample of individuals. These data were

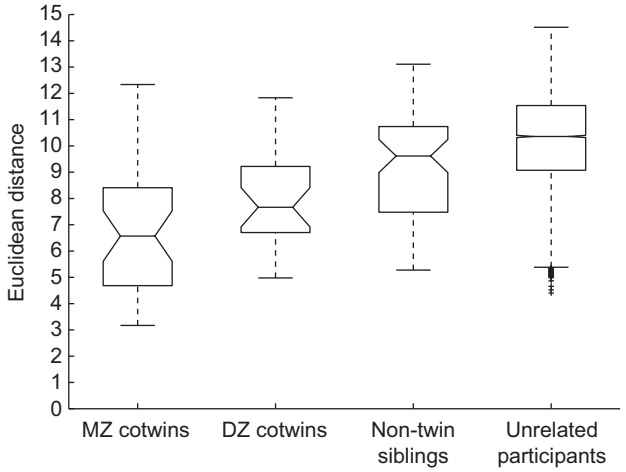

Figure 2 Box-whisker plots showing distributions of Euclidean distances between MZ cotwins ($N=37$ distances), between DZ cotwins ($N=28$ distances), among sex-matched nontwin siblings ($N=66$ distances), and among sex-matched unrelated participants ($N=8203$ distances) in the combined equated B1–B2 data set. The observations indicated with a plus sign in case of the unrelated participants illustrate the slight skewness of the distribution of the Euclidean distances among all participants.

Table 2 Significance of differences in Euclidean distances among subgroups[a]

	MZ cotwins	*DZ cotwins*	*Nontwin siblings*	*Unrelated participants*
MZ cotwins	–	–	–	–
DZ cotwins	>0.05	–	–	–
Nontwin siblings	<0.01**	>0.05	–	–
Unrelated participants	<0.01**	<0.01**	<0.01**	–

[a]*P*-values as resulting from multiple comparison test for differences in median Euclidean distances between monozygotic (MZ) cotwins, dizygotic (DZ) cotwins, sex-matched nontwin siblings, and sex-matched unrelated participants.
**$P<0.01$

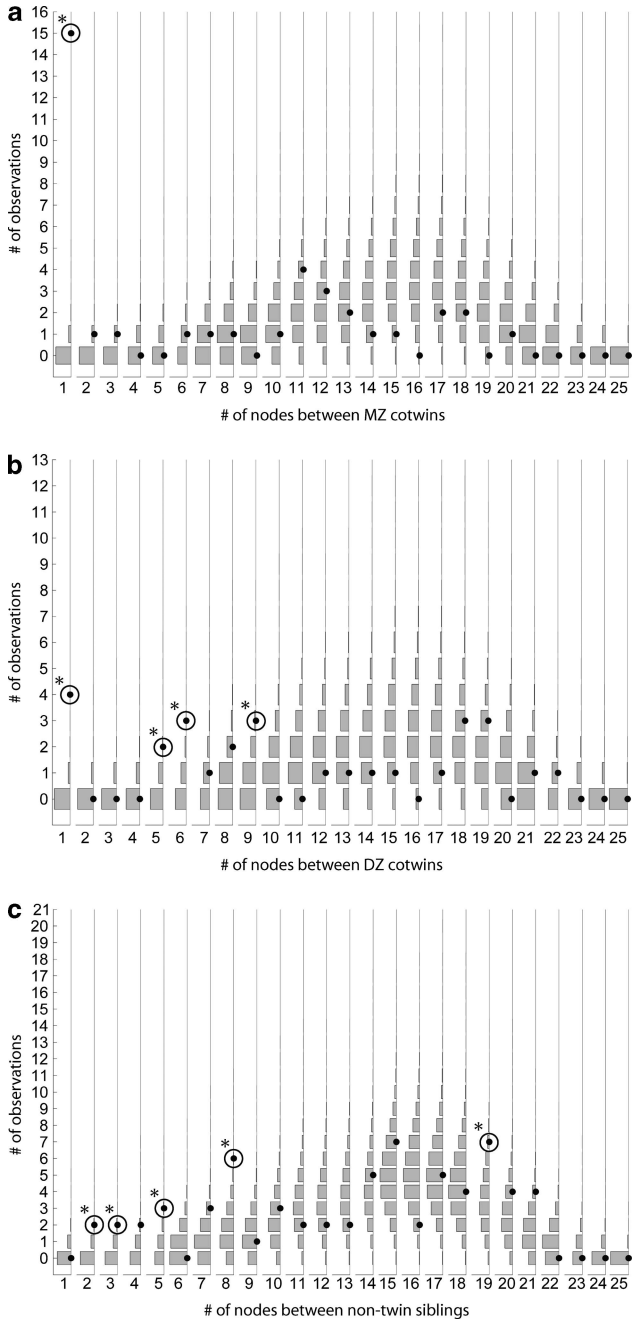

Figure 3 Results of node analyses with respect to permutation-based chance distributions. (**a**) MZ cotwins; (**b**) DZ cotwins; (**c**) Sex-matched nontwin siblings. Numbers of nodes separating cotwins or nontwin siblings increase from left to right in each panel. For each number of branching points, 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 1b and Supplementary Figure S2. The depicted chance distributions were created by combination of the results from all (ie, 100) sets of 10 000 permutations. Asterisks indicate average *P*-values <0.05 (see Supplementary Table S2).

generated with a metabolomics platform that analyzed lipids in blood plasma samples from twins and their biological nontwin siblings. Our results suggest that shared genetic background and shared environmental exposure contribute to similarity of blood plasma lipidomics profiles among individuals. First, we observed that median Euclidean distances, computed on the basis of these lipid profiles, were smaller between individuals with higher degrees of shared genetic background and/or shared environmental exposure. Second, when analyzing the numbers of nodes separating pairs of relatives in the hierarchical clustering dendrogram, in general relatives sharing relatively much genetic background and/or environmental exposure were separated by relatively small numbers of nodes. For example, the median intrapair Euclidean distances were smaller between MZ twins compared with DZ twins. We also observed a relatively small number of DZ twin pairs separated by only one node compared with the number of MZ twin pairs separated by one node. In addition, we found more DZ twin pairs separated by more than one node than was expected on the basis of chance. These observations together suggest that the higher proportion of genetic variation shared by MZ cotwins compared with DZ cotwins contributes to higher relative similarities of lipidomics profiles within MZ twin pairs.

The median Euclidean distance among biological nontwin siblings assumed a middle ground between the median distance between DZ cotwins and the median distance among unrelated participants. Although DZ cotwins and nontwin siblings share the same degree of genetic variation, the variation in age was larger among the nontwin siblings (range, 12–35 years) compared with the DZ cotwins (Table 1). This larger variation in age might have been one of the

most prominent factors contributing to the larger Euclidean distances among nontwin siblings compared with the DZ cotwins. The larger Euclidean distances among unrelated participants compared with nontwin siblings are as expected on the basis of the larger degrees of shared genetic and environmental variation among the siblings.

As we investigated a relatively young sample, almost all participants from the same family still lived together (with the exception of the older non-twin siblings). This may have enhanced the contrasts in genetic and environmental influences on phenotypic variation that we found for the different groups of participants: MZ twins, DZ twins, nontwin siblings, and unrelated participants. In a similarly composed sample of older individuals, for example, one might find a smaller influence of shared environment compared with our sample, as well as a larger relative influence of non-shared environment for all groups of participants. This would then result in more uniform results across the four types of participants in this older sample compared with our study sample.

It is conceivable that the genetic and environmental effects on lipid profile similarity in this study were mediated in particular via the TGs, because the lipids within this class comprised the largest number over all classes and displayed the highest correlation (median ρ, 0.74; median absolute deviation, 0.15).

Interestingly, we did not observe substantial changes in the clustering when expressing the concentrations of the reported TGs as a fraction of the total TG pool, suggesting that the levels of the individual TGs are driven by additional genetic and environmental factors other than those that determine the size of the VLDL-remnant fraction of human blood plasma.

Successful data set combination

We used two combined data sets, originating from two 'blocks' of measurements using the same LC–MS method. To correct the data for practically unavoidable nonbiological differences between the blocks, we applied the method of 'quantile equating' as described in one of our previous publications.[16] Although we showed in that paper that the correction by this method was successful at the metabolite level, our results in the current paper in addition suggest that after application of this method the clustering of family members is largely retained.

First, nine MZ twin pairs in the combined B1–B2 data of which the cotwins were separated by only one node, came from B1. All of these pairs were separated by only one node in the analysis of the separate B1 data as well;[14] in this analysis of the separate B1 data the total number of MZ twin pairs separated by one node was 13. The remaining six pairs of MZ cotwins 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 (results available by request); in the analysis of the B2 data separately there was only one additional pair of MZ cotwins separated by one node. However, another pair of MZ twins (belonging to the family with identifier '43', see the legend to Supplementary Figure S2) 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 latter observation suggests that due to quantile equating, the lipid profiles of the members of this particular MZ pair have been made more similar. This could be regarded a potentially unfavorable side effect of the data transformation effectuated by quantile equating, the primary aim of which is to remove block effects.

A second indication for successful block effect correction is that block effects were not significant in regression analysis of the number of nodes separating sex-matched relatives, both in the total group of participants, as well as in the MZ twins only.

Sex effects on clustering

We substantiate the finding that male MZ twin pairs have relatively more similar lipid profiles compared with female MZ pairs. In line with our previous results,[14] in linear regression we observed a significant association of female gender with an increased number of nodes in the dendrogram separating MZ cotwins. This observation might be related to asynchronous menstrual cycles (Supplementary Tables S3 and S4).

However, we could not replicate the almost perfect segregation of participants according to gender in the dendrogram as reported earlier.[14,28] In the dendrograms based on the separate B2 data set (data not shown), as well as in the combined B1–B2 data sets both before (Figure 1a) and after (Figure 1b) equating, we did not observe strong clustering of male and female participants. This might have been caused by the fact that in B1, for each study sample two replicate lipidomics analyses were performed, whereas in B2 each sample was measured only once. It is conceivable that the averaged replicate measurements from B1 approximate the true biological effects, for example, male–female differences, with higher precision than the single replicate measurements from B2.

Comparison with other studies

We are aware of two studies by other authors that allow for a comparison with our results obtained in the current study. Assfalg *et al*[28] provided evidence of specific, but not necessarily unique metabolic phenotypes in humans. The authors extracted an invariant part of the urinary metabolic phenotype as assessed by NMR analysis of multiple longitudinally (over a period of about 3 months) obtained samples. The results described in the current paper are concordant with those obtained by Assfalg *et al*, in the sense that healthy individuals appear to have a specific metabolic phenotype corresponding to a specific position of that individual in the multivariate 'metabolite space' put up by the metabolite concentrations as can be measured in for example biofluids.

Bernini *et al*,[27] on the basis of urine samples collected repeatedly in a cohort of 31 healthy individuals over a period of about 3 years, found that intra-individual differences in human metabolic phenotypes correlated mainly with changes in lifestyle, and were possibly also due to changes in microflora. In line with our results, the authors found that the members of the only MZ twin pair included in their study resembled each other more than any other pair of study participants did. They also observed that the second closest pair of individuals was a pair of father and son, and that in one individual metabolic similarities persisted despite significant changes in environment due to the moving to a different country. In combination, in accordance with our conclusions on the basis of the current study, these observations suggest an important role for genes in determining metabolic similarities among individuals.

Unique contributions of this study

In the current study, we build upon the work described in two of our previous publications to make some novel observations that are relevant for quantitative genetic analysis of lipidomics profiles.

First, due to the combination of the lipidomics data obtained in two different measurement blocks, we were able to include in this study many more DZ twin pairs and nontwin siblings compared with our previous publication that was based mainly on MZ twin pairs.[14] The additional participants in the current study enabled us to further establish the effects of genes, environment and age on relative similarities of lipid profiles among individuals.

Second, in this paper, we extend our hierarchical clustering approach by incorporating a number of statistical tests for the

observed effects. In particular, we tested the effects of monozygosity, CRP levels, gender and measurement block on the lipid profile similarity of sex-matched family members. Also, we report the results of tests for the comparison of the observed numbers of nodes in the dendrogram separating pairs of relatives, with the number of nodes as expected on the basis of permutation.

Furthermore, the results in the present report are consistent with successful correction of measurement block effects by 'quantile equating',[16] to make combinable metabolomics data sets for subsequent quantitative genetic analyses. This is a valuable addition to the publication where we introduced this data correction method, because the statistical tests in that paper mainly demonstrated the beneficial effects of quantile equating on the comparability of data for the same metabolites in different data sets. The current analyses support that this correction was indeed beneficial at the subtle level of the clustering of family members, which is informative of the relative contribution of genes and environment to phenotypic variation.

CONCLUSIONS

Taken together, our findings support the notion that shared genetic background and/or shared environmental exposures, as well as similarities in age, contribute to similarities in blood plasma lipidomics profiles among individuals. Furthermore, the results obtained in this study suggest that quantile equating is useful to make metabolomics data sets combinable for enhanced power in quantitative genetic analyses.

Application of our approach to genetically determined metabolite ratios, as identified by for example recent genome-wide association studies for metabolomics,[9,29,30] might provide even more distinct clustering of relatives than we have demonstrated here.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

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 program BioRange (project number: SP 3.3.1); the Netherlands Metabolomics Centre; Spinozapremie NWO/SPI 56-464-14192; the Center for Medical Systems Biology (CMSB); Twin family database for behavior genetics and genomics studies (NWO-MaGW 480-04-004) and NWO-MaGW Vervangingsstudie (NWO no. 400-05-717).

1 Mackay TF: The genetic architecture of quantitative traits. *Annu Rev Genet* 2001; **35**: 303–339.
2 Visscher PM, Hill WG, Wray NR: Heritability in the genomics era – concepts and misconceptions. *Nat Rev Genet* 2008; **9**: 255–266.
3 Eaves L: Putting the 'human' back in genetics: modeling the extended kinships of twins. *Twin Res Hum Genet* 2009; **12**: 1–7.
4 Sung J, Cho SI, Song YM *et al*: Do we need more twin studies? The Healthy Twin Study, Korea. *Int J Epidemiol* 2006; **35**: 488–490.
5 Posthuma D, Beem AL, de Geus E *et al*: Theory and practice in quantitative genetics. *Twin Research* 2003; **6**: 361–376.
6 Neale MC, Maes HM: *Methodology for Genetic Studies of Twins and Families*, Vol 67:KIuwer Academic Publishers: Dordrecht, 2005.
7 Gottesman II, Gould TD: The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* 2003; **160**: 636–645.
8 Comuzzie AG, Funahashi T, Sonnenberg G *et al*: The genetic basis of plasma variation in adiponectin, a global endophenotype for obesity and the metabolic syndrome. *J Clin Endocrinol Metab* 2001; **86**: 4321–4325.
9 Gieger C, Geistlinger L, Altmaier E *et al*: Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* 2008; **4**: e1000282.
10 Fiehn O: Metabolomics–the link between genotypes and phenotypes. *Plant Mol Biol* 2002; **48**: 155–171.
11 Shah SH, Hauser ER, Bain JR *et al*: High heritability of metabolomic profiles in families burdened with premature cardiovascular disease. *Mol Syst Biol* 2009; **5**: 258.
12 Kotronen A, Velagapudi VR, Yetukuri L *et al*: Serum saturated fatty acids containing triacylglycerols are better markers of insulin resistance than total serum triacylglycerol concentrations. *Diabetologia* 2009; **52**: 684–690.
13 Kriegel HP, Kröger P, Zimek A: Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transact knowl discov data* 2009; **3**: 1–58.
14 Draisma HHM, Reijmers TH, Bobeldijk-Pastorova I *et al*: Similarities and differences in lipidomics profiles among healthy monozygotic twin pairs. *OMICS* 2008; **12**: 17–31.
15 Hu C, van der Heijden R, Wang M, van der Greef J, Hankemeier T, Xu G: Analytical strategies in lipidomics and applications in disease biomarker discovery. *J Chromatogr B* 2009; **877**: 2836–2846.
16 Draisma HHM, Reijmers TH, van der Kloet F *et al*: Equating, or correction for between-block effects with application to body fluid LC-MS and NMR metabolomics data sets. *Anal Chem* 2010; **82**: 1039–1046.
17 Nederlands Tweelingen Register. http://www.tweelingenregister.org/, accessed 30 January 2012.
18 Willemsen G, de Geus EJ, Bartels M *et al*: The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* 2010; **13**: 231–245.
19 Barnes RJ, Dhanoa MS, Lister SJ: Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl spectrosc* 1989; **43**: 772–777.
20 Young G, Householder AS: Discussion of a set of points in terms of their mutual distances. *Psychometrika* 1938; **3**: 19–22.
21 Sokal RR, Rohlf FJ: The comparison of dendrograms by objective methods. *Taxon* 1962; **11**: 33–40.
22 R Development Core Team: R: *A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria 2009 ISBN 3-900051-07-0.
23 Kruskal WH, Wallis WA: Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 1952; **47**: 583–621.
24 Hochberg Y, Tamhane AC: *Multiple Comparison Procedures*. John Wiley & Sons: New York, 1987.
25 Suzuki R, Shimodaira H: Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006; **22**: 1540–1542.
26 Omori-Inoue M, Fukata H, Komiyama M *et al*: The contamination levels of organochlorines and the pattern of gene expressions in human umbilical cords from intra-pairs of twins at delivery. *Reprod Toxicol* 2007; **23**: 283–289.
27 Bernini P, Bertini I, Luchinat C *et al*: Individual human phenotypes in metabolic space and time. *J Proteome Res* 2009; **8**: 4264–4271.
28 Assfalg M, Bertini I, Colangiuli D *et al*: Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci USA* 2008; **105**: 1420–1424.
29 Illig T, Gieger C, Zhai G *et al*: A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* 2010; **42**: 137–141.
30 Suhre K, Shin SY, Petersen AK *et al*: Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 2011; **477**: 54–60.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)