'*Treelet Transform* for untargeted metabolomics data':

Treelet Transform generates serum metabolite and lipid components that are correlated to anthropometry and intestinal microbiota in a cross-sectional EPIC-Potsdam sub-study

vorgelegt von

Diplom-Biologin,

Master of Science in Epidemiology

Jana Förster

von der Fakultät VII – Wirtschaft und Management der Technischen Universität Berlin

Dissertation zur Erlangung des akademischen Grades

Doktorin der Gesundheitswissenschaften/Public Health

-Dr.P.H.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzende: Prof. Jacqueline Müller-Nordhorn

1. Gutachter: Prof. Dr. Heiner Boeing

2. Gutachter: Prof. Dr. Reinhard Busse

Tag der wissenschaftlichen Aussprache: 05. August 2014

Berlin 2014

Table of contents

List of figures

List of applied formulae

List of tables

List of supplemental tables

List of abbreviations

BAI	Body adiposity index
BMI	Body mass index
CRP	C-reactive protein
DGGE	Denaturing gradient gel electrophoresis
DNA	Deoxyribonucleic acid
EPIC	European Investigation into Cancer and Nutrition
GC x GC-TOFMS	Two-dimensional gas chromatography coupled to time-of-flight mass spectrometry
GFPQ	German food preference questionnaire
HbA1c	Glycosylated haemoglobin, Type A1C
MRT	Magnetic resonance tomography
Nb	Number
NCD	Non-communicable diseases
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
PC	Phosphatidylcholine
PCR	Polymerase Chain Reaction
PE	Phosphatidylethanolamine
rRNA	Ribosomal ribonucleic acid
Spmets	Small polar metabolites
TC	Treelet component
TG	Triglyceride
TT	*Treelet Transform*
UPLC	Ultra-performance liquid chromatography
WHO	World Health Organization
WHR	Waist-to-hip ratio
WHtR	Waist-to-height ratio

1. Introduction

1.1 Background

Controlling the escalating epidemic of overweight and obesity is according to the World Health Organisation (WHO) one of the most obvious public health problems not only in developed parts of the world but also in developing countries [1]. Overweight and obesity already start to develop in childhood [2] and are in many cases attributes of a Western lifestyle, which is characterised by higher energy intake compared to the energy expenditure and so by an imbalance of energy intake and expenditure [3]. Further both, obesity and overweight, or even the causative behavioural aspects of poor diet and physical inactivity, are among others the major risk factors for non-communicable diseases (NCD), e.g. cardiovascular diseases, type 2 diabetes and cancer [4-6], accounting for the highest mortality worldwide [7]. On the one hand, prospective cohort studies showed a clear association between anthropometric parameters indicating obesity such as an elevated waist circumference and disease risk and even with the risk of death [8]. On the other hand, further studies revealed single metabolites as risk factors for obesity [9] and also for type 2 diabetes, coronary events and cancer [10, 11]. It is useful to figure out valid biomarkers for overweight and obesity in order to identify individuals with an elevated risk prematurely since most of the metabolites -and together with them the deleterious anthropometrical traits- are influenced by diet [12] and lifestyle factors [13] and by that are modifiable. For the modification of such deleterious traits individualised prevention strategies can be developed, but it is necessary to characterise and identify persons that are exposed to a deleterious anthropometry and also have a metabolic profile that is combined with the anthropometric parameters and increases risk for NCD.

Recent developments in biochemical techniques for metabolite determination enable the assessment of not only a few, pre-defined metabolites [14]. Chromatography, mass spectrometry (FIGURE 1), nuclear magnetic resonance (NMR) and other methods allow the investigation of hundreds of compounds or even all detectable compounds in one biological sample [15].

These comprehensive techniques set the stage for a new research field called ‘metabonomics’, or the more common ‘metabolomics’, at the end of the last century [16]. Metabolomics aims to characterise and quantify up to all small molecules in a sample [17, 18] and by that is more appropriate to assess the complexity of the human metabolism than investigations on single metabolites.

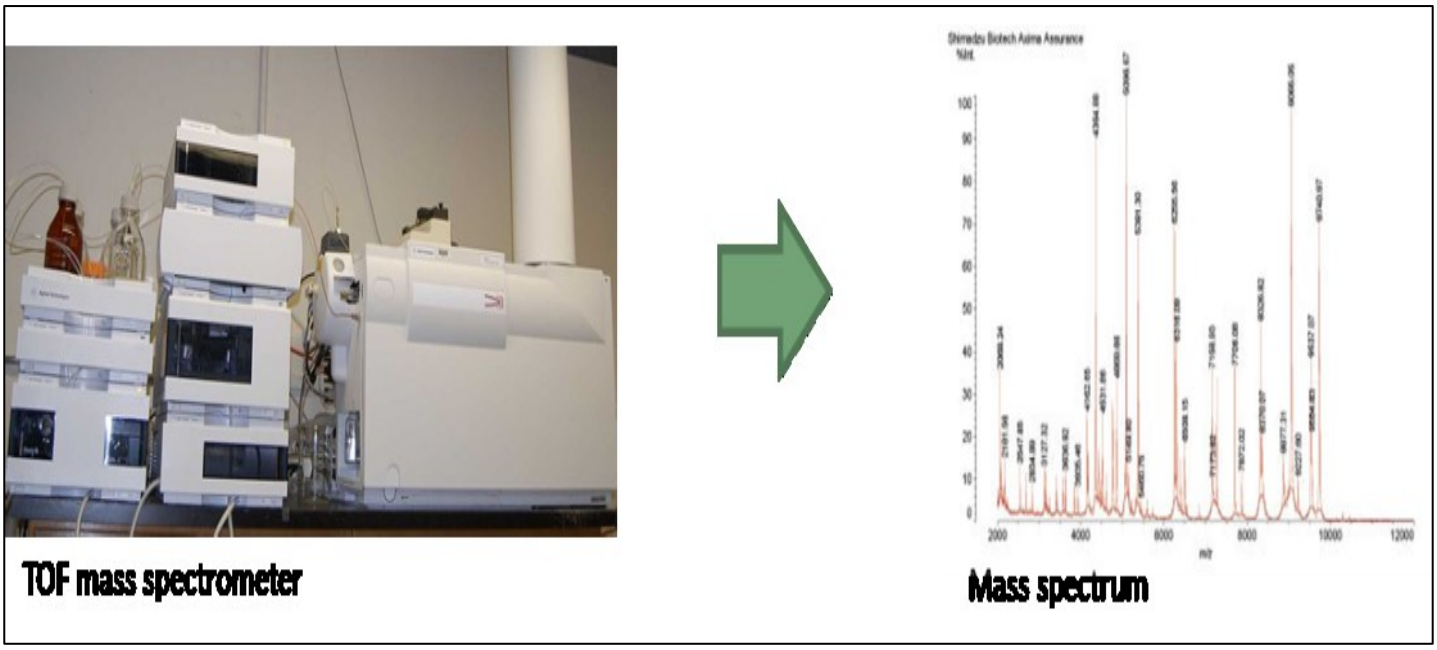


FIGURE 1: Time-of-flight mass spectrometer and a mass spectrum; image sources: http://upload.wikimedia.org/wikipedia/commons/e/e2/ESI_TOF.jpg and http://www.shimadzu.de/axima-assurance

First metabolomic studies pursued a targeted approach, quantifying a pre-defined set of metabolites in human tissue. Such studies found an association between amino acids and phosphatidylcholines and type 2 diabetes [19] as well as between plasma choline concentrations and the risk for cardiovascular diseases [20]. A next step in the metabolomics research is the utilisation of untargeted approaches aiming at the global detection of all, including yet unknown, metabolites in a sample.

Applying metabolomics, in particular untargeted metabolomics, in epidemiological studies is accompanied by some challenges. There are the aspects of biological and technical variation, respectively. The measurement of hundreds of compounds results in a higher measurement error than determining only a single metabolite. Furthermore, a high biological variation in terms of inter-individual as well as within-person variation is assumed but not yet clarified. Both aspects can affect the results of a study and have to be considered in the interpretation [21].

Another challenge that comes along with metabolomics research is the huge amount of inter-correlated data that is generated for a limited number of subjects due to high laboratory effort. Most of the metabolites are not independent, but related to each other, a fact that has to be considered in subsequent analyses. In the case of correlated data associative studies of single metabolites and an outcome of interest would disregard the data structure and will lead to unstable and unsatisfying results. The frequent constellation of much more variables (p) than observations (n), is obstructive for data reduction procedures [22] that summarise correlated variables in latent variables and are advisable in advance to associative analyses.

There are several established methods for dimension reduction of complex data, such as principal component analysis (PCA), cluster analysis or reduced rank regression [23, 24], but they often generate unstable results, that are very complex and difficult to interpret [22]. A method reducing dimension in correlated data and resulting in sparse latent variables that are more simply to interpret is the *treelet transform* (TT) procedure. TT is based on a treelet algorithm grouping together the most correlated variables and generating a sum variable, which replaces the correlated variables, and a residual variable [22, 25]. TT combines a hierarchical cluster analysis and local PCA producing components containing subsets of the original variables [26, 27].

In the present study, data on serum small polar metabolites and serum lipids, that were shown to be associated to NCD and mortality [19, 28, 29], are investigated regarding their correlation structure. Applying TT as an exploratory approaches summarising, latent components are generated. The components are compared to factors resulting from a PCA on the same data set to evaluate the TT results critically and compare TT to an established dimension reduction method. The obtained components and factors will be associated to anthropometric parameters, such as body mass index (BMI), waist circumference and body fat parameters, measured and obtained by an algorithm on MRT data, and to intestinal microbiota data. The investigation of the relation of metabolites and intestinal microbiota was chosen because former studies showed an influence of microbiota on human metabolites [30-32].

These analyses will contribute to get a better idea on how serum metabolites and lipids are associated to anthropometry and intestinal microbiota. The associations are regularly found to exist in different studies [33, 34].

1.2 Public health relevance

Against the background of an ageing population and increasing prevalence of NCD the early recognition and prevention of risk factors become more and more important. Proceedings in medical research enable a better medication of chronic diseases resulting in a higher life expectancy of diseased. Additionally, generally better life circumstances and higher life expectancies cause increasing costs and depict a challenge for the public health system. Adiposity negatively influences life quality [35] and it is inversely related to healthy years of life [36]. According to the WHO in 2008, more than 1.4 billion adults have been overweight worldwide, meaning they have a BMI of at least 25kg/m². Out of them about 500 million have a BMI of at least 30 kg/m², what is defined as obese. Within 30 years, from the 1980s up to now, obesity rates nearly doubled [37] and in specific groups the obesity rates are extremely high, e.g. in persons with mental ill-health almost 80% are overweight or even obese in an American study population [38, 39]. In 2012, the German Robert Koch Institute reported that nearly 25% of the German population aged over 18 years are obese [40]. In Germany health costs provoked by adiposity increased by 10% to 863 billion euro in 2008 compared to 2002 [41] and it is assumed that in the UK the costs will double until 2050 [39].

In worldwide studies overweight and obesity were also identified as a public health challenge with an increasing trend due to the recent trend that more and more people become overweight [42].

Except for some factors, such as genetic predisposition [43, 44] or metabolic dysfunctions [45], causes of obesity are of behavioural nature. In most cases the circumstance of a higher energy intake than expenditure, an impaired energy balance, resulting from a high caloric diet and physical inactivity [46], gives frequently rise to overweight and obesity. The circumstance that most of the risk factors are behavioural and thus are modifiable offers the chance to actively counteract the development of obesity.

To apply effective prevention strategies that work on the right individuals it is necessary to identify persons under a high risk for obesity as early as possible and to better understand underlying biological pathways in the development of overweight and obesity. To accomplish these aims studies investigating the relationship between metabolites and measures of obesity are required and could contribute to improved prevention strategies of NCD.

1.3 Metabolomics in epidemiology

Investigation of human metabolome is of special interest in research on human health as it helps to clarify pathogenic pathways. Starting with studies on single biomarkers and risk factors, epidemiological research evolved from classic into system epidemiology considering the whole complex of the human metabolism (FIGURE 2). Advancements in the '-omics' sciences enable a better investigation of biological processes by exploring the overall system rather than single factors within the human metabolism. In the order of their occurrence in a metabolic pathway genomics, transcriptomics, proteomics and finally metabolomics explore complex networks rather than single genes, transcripts, enzymes or products of biological reactions. Mostly these approaches have been applied in animal studies or small clinical trials and not on a population base. Combining the complex approach of metabolomics and the efficient setting of observational prospective cohort studies constitutes a real chance to enlighten mechanisms and pathways underlying associations that have been observed between environmental influences and disease risk (FIGURE 2) in epidemiological studies for several years [47].

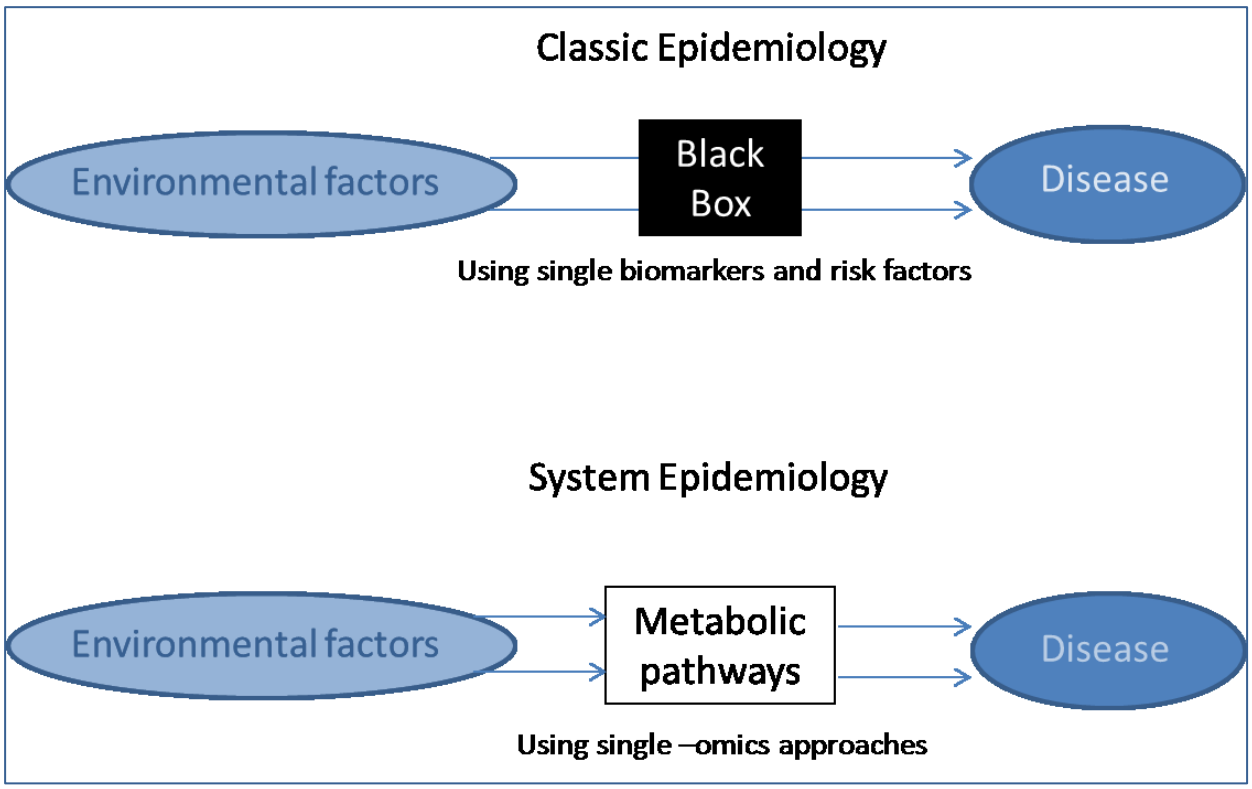


FIGURE 2: Classic 'black-box' epidemiology vs. system epidemiology using metabolomics to illuminate metabolic pathways; figure is of author's own devising

To investigate assumed mechanisms with some prior knowledge on involved compounds, in order to verify and extend existing hypotheses, targeted metabolomics can be applied. This bottom-up method precisely quantifies a predefined set of up to some hundred metabolites in a sample. In contrast to targeted metabolomics the concept of untargeted metabolomics has a more exploratory nature being a top-down method, which determines all substances of various chemical classes present in a sample (TABLE 1).

TABLE 1: Characteristics of targeted and untargeted metabolomic approaches; table is of author's own devising

	Targeted metabolomics	Untargeted metabolomics
Aim	Determine a pre-defined set of metabolites in a sample → hypothesis driven	Determine globally all detectable metabolites in a sample → hypothesis generating
Methods	- Chromatographic and mass spectrometric approaches - Quantification with standard curves or comparison with standard groups	- Chromatographic and mass spectrometric approaches - Nuclear magnetic resonance - Analysis of mass spectra with the help of databases for identification and quantification of detected metabolites
Result	Data set with concentrations of pre-defined set of metabolites → quantification of specific metabolites	Data set with concentrations of all existing metabolites, identified and unknown ones → global metabolic profile
Advantages	- Precise quantification of metabolites	- Determination of a high number of metabolites in many samples - Approximation on real existing metabolic pathways - Small sample amount necessary
Disadvantages	- Determination of a limited number of metabolites - Potential disregard of true metabolic pathway	- Manual - Time-intensive - Limited data quality, high variance

The number of determined metabolites attains several thousands and by that is much higher than by the use of a targeted approach. This situation enables a more comprehensive research regarding metabolic processes which promote disease development but it also challenges statistical methods to achieve stable results that allow simple interpretation.

Many biochemical molecule quantification methods, such as mass spectrometry or nuclear magnetic resonance spectroscopy have been developed during the last years and enable the detection of numerous compounds in biological samples. This is a development, which may help to achieve better insights into the pathogenesis of common diseases [28, 48].

1.4 Methods for the processing of high-dimensional data

The achievements in laboratory measurements enable an assessment of high quality data in epidemiological studies. But the high-dimension setting where the number of variable is much larger than the sample size challenges the scientists. There are a lot of established methods evaluated regarding their possibilities to deal with high-dimensional data [49]. However, it is necessary to develop new and to examine unestablished methods regarding their process of high-dimensional data. In many studies with more variables than observations either all assessed variables were taken into account and applied in analyses separately or for potentially confounding variables is adjusted. But this disregards the potential inter-correlation structure of the variables. This is considered with statistical methods summarising these variables that are correlated to a latent variable, such as factor analysis, PCA and cluster analysis. Some of these methods generate latent variables that can be used for further analyses, but have to be interpreted regarding the underlying original variables. This again is a challenge, since for example components resulting from a PCA are formed by all original variables. Further, a cluster analysis generates clusters of strongly correlated original variables, which enable a clear interpretation, but it generates no summarising latent variables that can be used in further analyses. Further cluster analysis does not consider the inter-correlation of the original variables but the correlations between the observations [50].

These difficulties long for further methods that also regard to the inter-correlation structure but generate sparse latent variables. This approach is implemented with the method *treelet transform* (TT), it summarises variables regarding their inter-correlation structure and generates latent variables that combine a limited number of original variables.

1.5 Study questions and objectives

The present thesis aimed to investigate a comprehensive data set from a cohort study generated among others with an untargeted metabolomics approach. The inter-correlation structure of serum metabolites and lipids and its association to anthropometric parameters and intestinal microbiota were investigated. Thereby, different statistic approaches were used to resolve the following questions of interest:

- Is it possible to reduce dimension of serum metabolite and lipid data to only a few latent components with as little loss of information as possible with the approach of TT?
- Are the results comparable to results from a PCA on the same data set as an example of an established dimension reduction method?
- Are the retained TT components and PCA factors associated to anthropometric and microbial parameters and, if so, in the same way?

To clarify these questions, 121 serum metabolites and 353 serum lipids were summarised to treelet components and to factors from PCA, which explain portions of the variance within the original data. Components and factors were compared concerning structure and were correlated to the outcomes of interest.

The thesis aimed to evaluate the utility of the components resulting from TT, as an unestablished but encouraging method, within system epidemiology.

2. Methods

2.1 Study population

Participants of the present study have been selected randomly within the EPIC (European Prospective Investigation into Cancer and Nutrition)-Potsdam cohort. EPIC- Potsdam is a part of an Europe-wide cohort study including more than 500,000 participants in ten European countries [51] which is coordinated by the International Agency for Research on Cancer (IARC) and the World Health Organization (WHO). The overall aim of the study is to investigate the relationship between dietary habits, metabolic and genetic factors, and chronic diseases, especially cancer. The EPIC-Potsdam part of this study comprised about 27,500 participants that were aged mainly between 35 and 65 years in women and 40 to 65 years in men. The participants have been recruited between 1994 and 1998 in Potsdam and the surrounding area [52].

In the context of a sub-study to validate and calibrate data on physical activity, anthropometry and nutrition within the EPIC-Potsdam follow-up, 816 participants have been randomly selected from the initial study population. The study has been proven and permitted by the ethics commission of the State Medical Association of Brandenburg in Cottbus (Brandenburg, Germany).

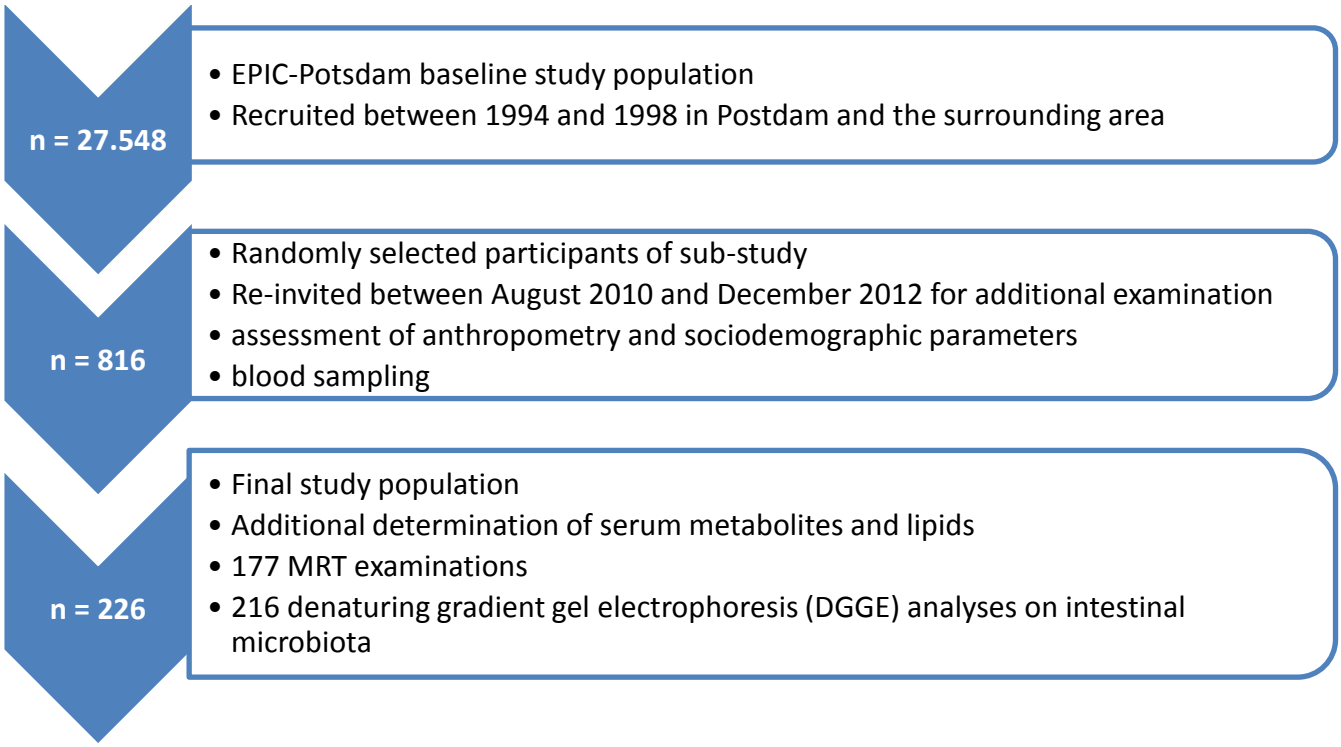


FIGURE 3: Study population for the present investigation

All participants provided a written informed consent before participating in the examination. 226 of these sub-study participants were included in the present investigation (FIGURE 3). These 226 were the first EPIC participants with a second blood sample and updated information on anthropometry.

2.2 Data assessment

Between August 2010 and December 2012, 816 EPIC-Potsdam participants have been re-invited for additional examinations including blood and stool sampling, anthropometry, assessment of physical activity (GPAQ and activity monitoring instrument: Actiheart), questionnaires and interviews on habitual diet (GFPQ and 24 hour dietary recalls) and further non-dietary aspects of life style (LEGU).

2.2.1 Blood sampling

Blood samples were drawn by a study physician following a standardised operating procedure. Whole blood was partly analysed the same day regarding clinical blood parameters (ALAT, AlbuSN, ASAT, Basophile, Bilirubin (direct and overall), blood sugar, cholesterol, creatinine, high-sensitivity C-reactive protein (CRP) (quantitative), eosinophiles, erythrocytes, gamma-GT, haemoglobin, HbA1c (% and mmol/mol), HDL cholesterol, haematocrit, uric acid, LDL cholesterol (chemical and calculated), leukocytes, lymphocytes, monocytes, neutrophiles, thrombocytes, triglycerides). Another part was separated in its fractions (serum, plasma, buffy coat and peripheral mononuclear cells) and stored in liquid nitrogen (-196°C) until further analyses.

For 226 participants the serum fractions were investigated regarding all existing compounds within the serum with an untargeted metabolomics approach. Serum metabolites were determined with a two dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC x GC-TOFMS) and serum lipids with an ultra-performance liquid chromatography (UPLC) as described by the group of Orešič et al. [53]. The GC x GC-TOFMS method assessed various small polar metabolites (spmets) including amino acids, fatty acids and sugar compounds (FIGURE 4), whereas the UPLC assessed serum lipids, such as di- and triglycerides, ceramides, sphingomyelines, cholesteryl esters (FIGURE 5), phosphatidylcholines and -ethanolamines and lyso-phosphatidylcholines and -ethanolamines (FIGURE 6). Overall 1039 lipids and 587 spmets were detected including 392 and 134 identified compounds, respectively. To identify peaks from the mass spectra two criteria were used. First, the mass spectra were compared to commercial library spectra. If the match was higher than 850 (at a maximum match of 1000) this criterion was met. Second, corrected retention times were calculated and matched with literature values for the

potential compound, if the difference was less than 25 the second criterion was also met. If several spectra fulfilled both criteria, the most similar match was designated as this compound, whereas all other resembling spectra also got the same terminology but were additionally marked with a star (*) to identify them as similar but not completely matching compounds. For three fatty acids (arachidonic acid, linoleic acid and stearic acid), seven sugar compounds (d-fructose, d-galactose, d-glucose, erythrose, glucopyranose, ribonic acid and xylitol) and four other compounds (lactic acid, butanoic acid, serine and urea), there was more than one spectrum fulfilling the matching criteria for one compound present.

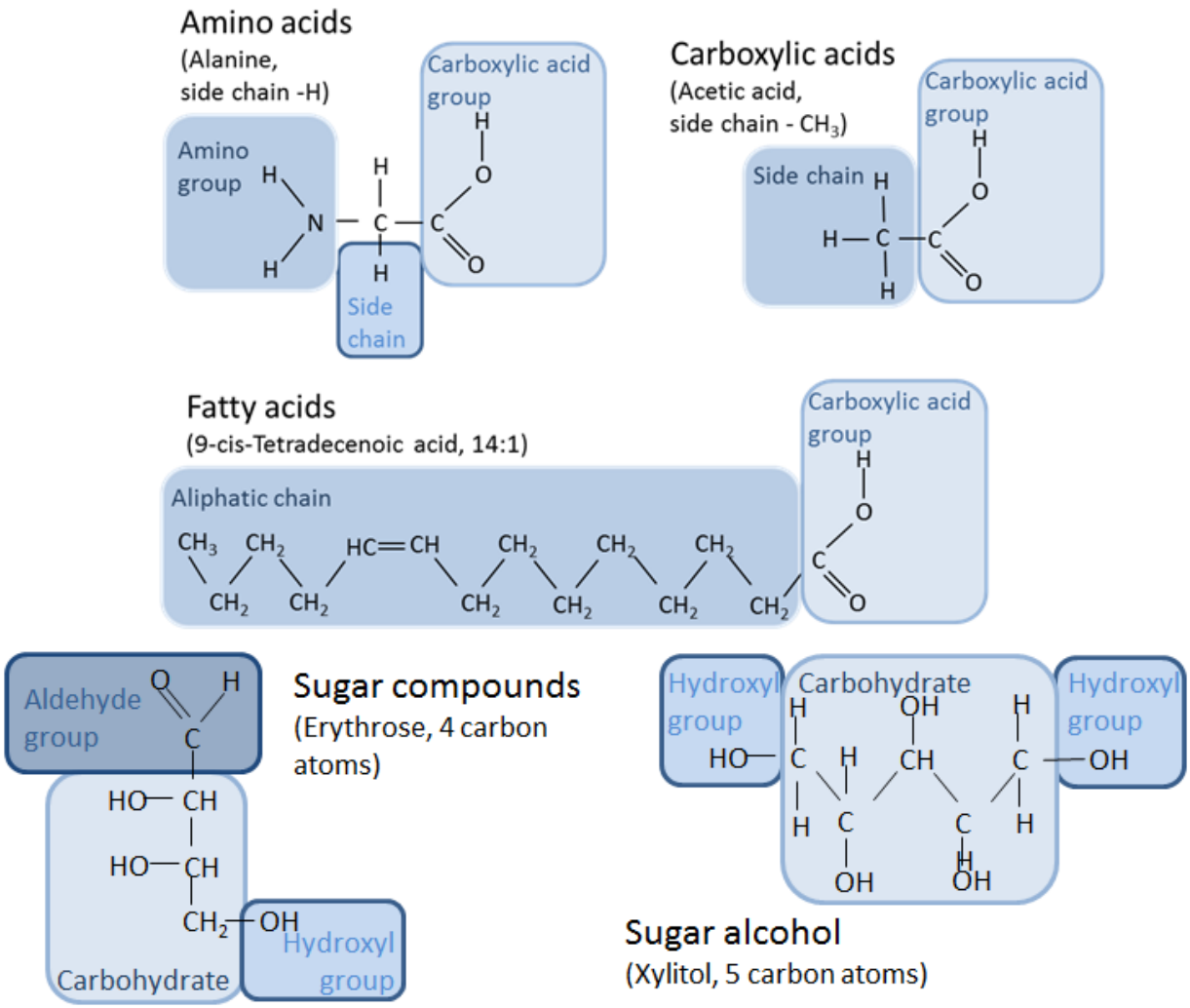


FIGURE 4: General structure of detected and identified small polar metabolites, i.e. Amino acids, carboxylic acids, fatty acids, sugar compounds and sugar alcohols; figure is of author's own devising

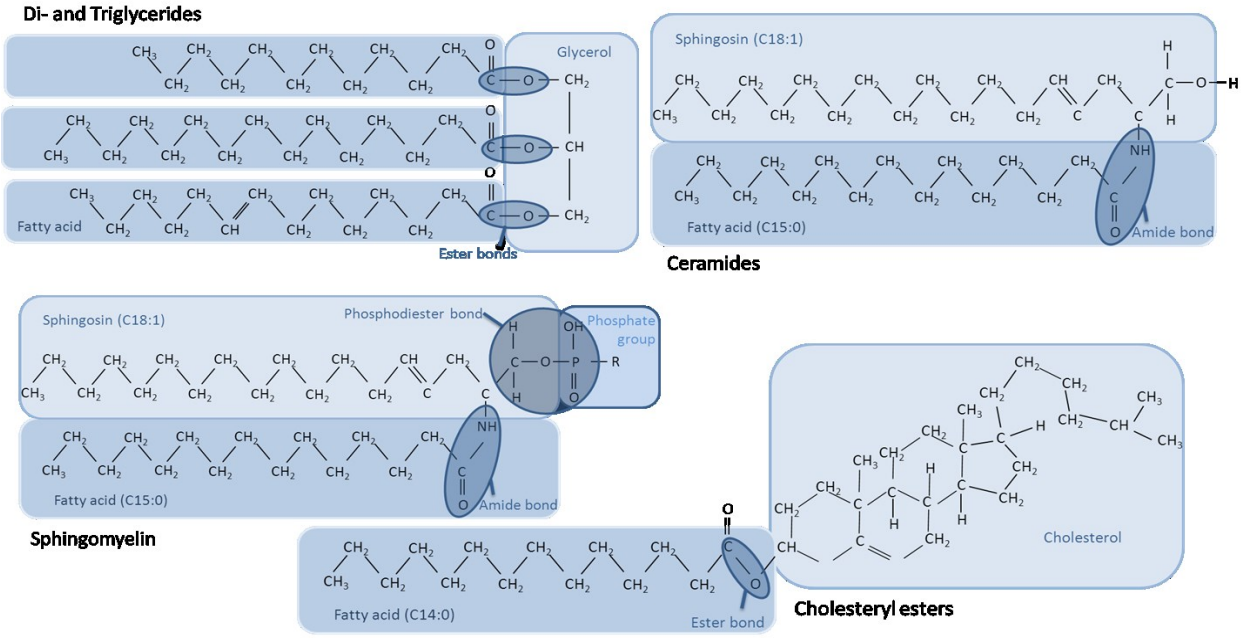


FIGURE 5: General structure of detected and identified di- and triglycerides, ceramides, sphingomyelins and cholesteryl esters; figure is of author's own devising

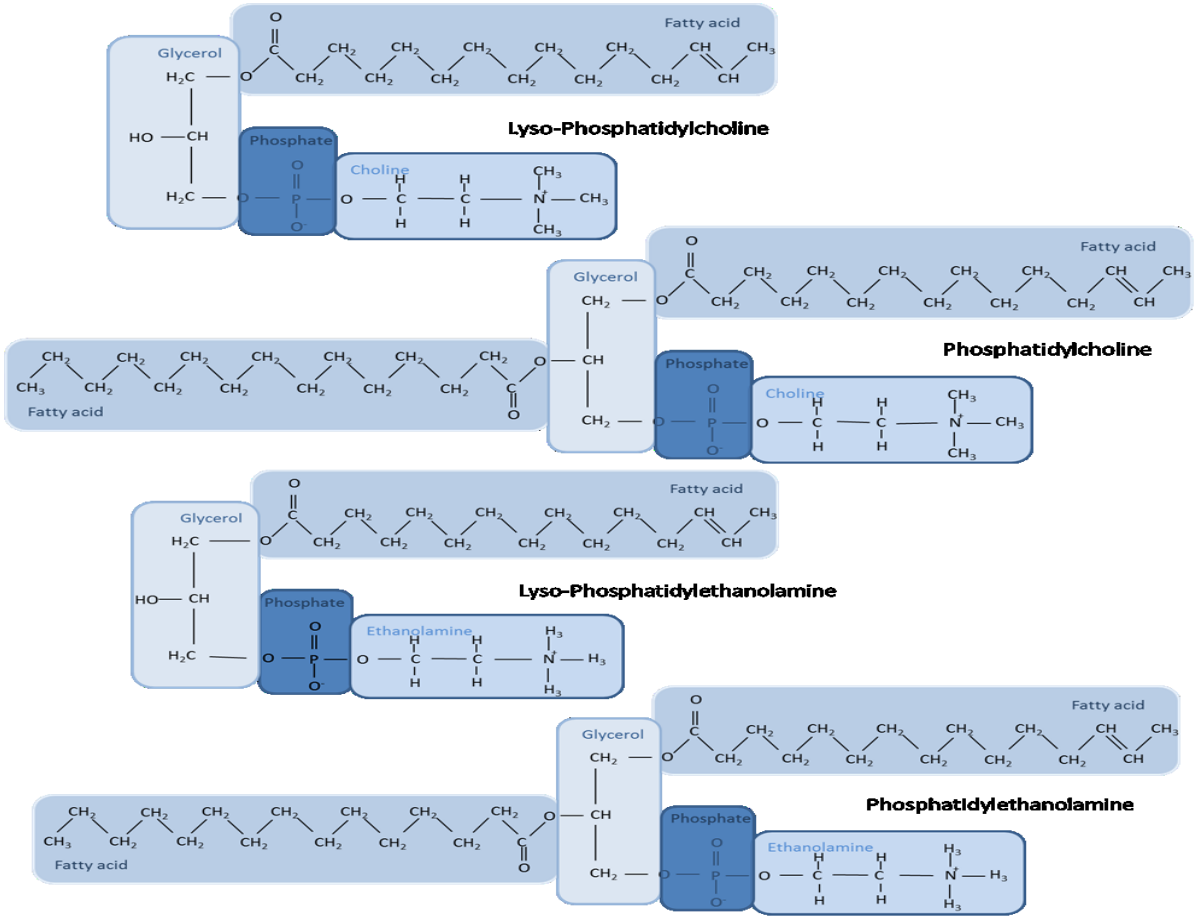


FIGURE 6: General structure of detected and identified lyso-phosphatidylcholines, phosphatidylcholines, lyso-phosphatidylethanolamines and phosphatidylethanolamines; figure is of author's own devising

2.2.2 Anthropometry

For elicitation of anthropometry the participants' body height, body weight, waist and hip circumferences and skin fold thickness of upper arm, upper and lower leg, waist and hip were measured. All parameters were measured by trained study personnel following the guidelines of the WHO [54] and a standard operating procedure.

Body height was measured by stadiometer (seca 217). For assessment of body weight an electronic personal scale (seca 876) was used. All circumferences were determined by using an inelastic measuring tape (seca 201, seca deutschland, Hamburg, Germany). A Lange skin fold calliper (Beta Technology, Santa Cruz, CA, USA) was used to measure thickness of skin folds. Thorax depth and width were determined by means of a thorax calliper (GPM, Zurich, Switzerland).

Body weight was measured in kg with an accuracy of 0.1kg. Body height, all circumferences, thorax depth and width were measured in cm and skin fold thickness was measured in mm. The measurement precision was 1mm for body height, skin fold thickness and thorax measures and 5mm for circumferences.

BMI was calculated as quotient of body weight in kg and body height in m squared. Waist-to-hip ratio (WHR) was calculated as quotient of waist circumference and hip circumference. Waist-to-height ratio (WHtR) was calculated as quotient of waist circumference and body height.

Body fat mass was calculated using formulae of Durnin and Womersley [55] taking into account thickness of skin folds on biceps, triceps, back and hip (FORMULA 1). The formulae are sex-specific, for men other permanent values are used than for women.

a)

$$Body\ fat\ mass^{women} = \left(4.95/\left(1156.7 - 71.7 * \log\left(\frac{triceps + biceps + back + hip}{1000}\right) - 4.5\right)\right) * 100$$

b)

$$Body\ fat\ mass^{men} = \left(4.95/\left(1176.5 - 74.4 * \log\left(\frac{triceps + biceps + back + hip}{1000}\right) - 4.5\right)\right) * 100$$

FORMULA 1: Formulae to calculate body fat mass (in percent) for women (a) and men (b) using triceps, biceps, back and hip skinfold thickness (in mm) according to Durnin and Womersley [55]

Further, the body adiposity index (BAI) was calculated to get an additional measure of body fat mass [56]. This index determines percentage of body fat amount by taking into account hip circumference and body height (FORMULA 2) without differentiating between men and women.

$$\text{BAI [\%]} = (\text{hip circumference [cm]}/(\text{body height [m]})^{1,5}) - 18$$

FORMULA 2: Formula to calculate percentage of body fat amount considering hip circumference and body height according to Bergman et al. [56]; BAI- body adiposity index

Additional markers for body fat mass were taken from MRT measurements which enable to calculate visceral and subcutaneous fat in cm^3. Most participants (n=177) underwent a MRT analysis which provided information on subcutaneous, visceral and total abdominal fat (in cm^3) by utilization of automatic quantification algorithms [57].

2.2.3 Intestinal microbiota

The participants provided a faecal sample for determination of intestinal microbiota composition using a 16S rRNA based polymerase chain reaction (PCR) in combination with denaturing gradient gel electrophoresis (DGGE). The sample was collected at the participant's home, was frozen immediately at -18°C and brought to the institute by collection of a driver from the institute within seven days without thawing where it was stored at -20°C until analyses. DNA was extracted using the FastDNA Spin Kit for Soil (MP Biomedicals) with a modified protocol [58] and V6-V8 region of bacterial 16S rRNA was amplified in a PCR with bacteria specific U968-GC-f and U1401-r primers [59]. PCR amplification was performed as follows: initial denaturation at 94°C for 5min, denaturation at 94°C for 30s, followed by 35 cycles with an annealing temperature of 50°C for 20s, and primer extension at 72°C for 40s, and a final extension for 7min at 72°C.

150ng of PCR products were analysed regarding their GC content by using the DCode Universal Mutation Detection System (BIORAD Laboratories, Hercules, California) and separating them in a DGGE to generate sample specific band patterns (FIGURE 7).

Band patterns were analysed using the Phoretix 1D and 1D Pro software package (TotalLab, Newcastle upon Tyne, UK) [60]. The analysis of the scans provided dichotomous information on the presence or absence of each potentially detectable DGGE band in each sample enabling the investigation of gut microbiota composition.

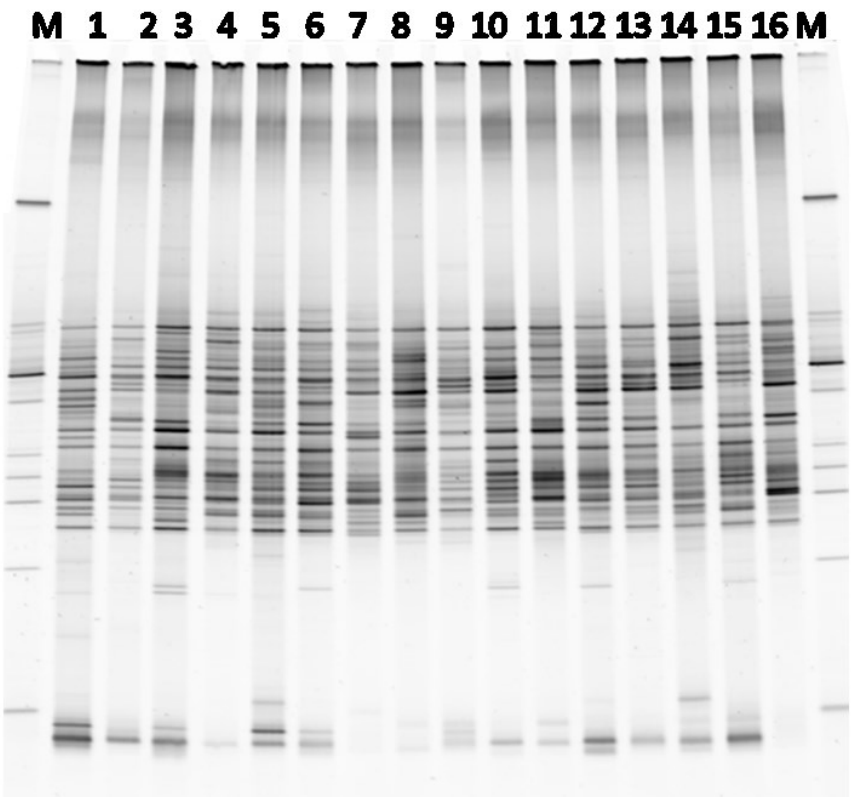


FIGURE 7: Sample specific DGGE band pattern of 16 samples, M – marker, numbers – sample numbers

Diversity among intestinal microbiota can be described on the one hand by the absolute number of different DGGE bands, which can be detected overall in one sample. On the other hand diversity can be expressed by the Shannon-Wiener (D_{SW}) (FORMULA 3) diversity index using the following formula

$$D_{sw} = -\ln (1/\text{total number of DGGE bands})$$

FORMULA 3: Shannon-Wiener diversity index

Other studies also use this index for the representation of the microbial diversity [61, 62].

In the present study the diversity described by the number of DGGE bands is referred to 'diversity 1' and the diversity described by the Shannon-Wiener index is abbreviated 'diversity 2'.

216 participants provided sufficient sample to receive information on intestinal microbiota, therefore 99 men and 117 women could be included into the analyses in question.

2.3 Statistical analyses

The present study aims to investigate the inter-correlation structure within a set of serum metabolite and serum lipid variables employing the dimension reduction method *treelet transform* (TT) resulting in latent treelet components (TCs). TCs were compared to factors from a PCA as an established dimension reduction method on the same data set (FIGURE 10). Associations of generated components and factors to anthropometric parameters were investigated with partial correlations. For TT analyses the STATA add-on "tt" [63] was used in STATA (STATA 10.0, StataCorp. LP, Texas, USA). All other statistical analyses were conducted using SAS (SAS enterprise guide version 4.3; SAS Institute Inc, Cary, NC, USA).

2.3.1 Descriptive analyses

The study population was characterised regarding relevant variables, such as age, sex, anthropometry, serum metabolites and lipids, and intestinal microbiota. Continuous variables were reported with mean and standard deviation, for categorical variables percentages were indicated.

2.3.2 Dimension reduction with *Treelet Transform (TT)*

Data on serum metabolites (p=587) and serum lipids (p=1039) were divided into identified metabolites (p=134) and lipids (p=592) and unidentified ones. In the present study only identified metabolites and lipids were analysed. TT is an exploratory, data-driven dimension reduction method combining cluster analysis and local PCA resulting in an appropriate number of latent variables - the TCs - explaining a large section of the variance within the investigated data [22, 26].

The flow chart in FIGURE 8 shows the principal strategy of a TT calculation in several steps. It starts with the definition of the cut-level and the number of TCs to extract and results in TCs and individual component scores.TT clusters all variables regarding their correlations in a hierarchical manner. In a first step it clusters the two most correlated variables out of all variables (p) and runs a local PCA on these two variables. This step generates two new variables – a sum variable with the largest variance and a residual variable with the lower variance. By keeping the sum variable and discarding the residual variable a new cluster analysis on p-1 variables is conducted. This procedure is repeated p-1 times until all variables are summarised in one cluster. This cluster can be visualized in a cluster dendrogram - a binary tree with p levels. The nearer two variables in this tree are the more they are correlated. To plot the dendrograms, distance matrix was exported from STATA and the *proc*

tree procedure of SAS was used. To obtain exact position of variables within the tree *lineprinter* option in *proc tree* was applied. Based on the dendrogram TCs are generated. The tree is cut on a predefined level considering aspects of complexity of TCs, proportion of explained variance and stability of results. Cross-validation methods can help to choose a suitable cut-level. Therefor the data set is split into equal sized subsets and in each subset a TT is conducted. For each possible cut-level the TCs with the highest variance are calculated in all but one subset. Within this one data set the sum of variances explained by the components obtained from all other subsets is calculated. This procedure is repeated several times resulting in a cross-validation score for each cut-level. A suitable cut-level is found when an increase of the cut-level does not substantially increase the cross-validation score [63]. Once a cut-level is chosen, the variance for all components at this level is calculated and a predefined number of TCs explaining the highest proportions of variance is retained. To specify the number of retained TCs it is reasonable to use a scree plot, which plots the explained variance against the number of components. The optimal number of retained components is when increasing the number of components does not substantially increase the proportion of explained variance. The retained TCs are loaded by all original variables constituting the cluster that the component depicts. All other original variables that are not in the cluster have no loading on the component.

To assess stability of the results a subsampling method is used. The TT is repeated many times (in most cases 100 times) in a subsample of 80% of the original data. From every TT the same number of TCs as in the original analysis is retained and the TCs are transformed into sign patterns in which the variables with a positive loading on the component get a “+”, that with a negative loading get a “–“ and that variables not loading on the component get a “0”. In a next step the relative frequencies of all sign patterns that occur in at least 10% of the subsampling repetitions are displayed. By this it can be evaluated how often the retained TCs occur in a subsample and how stable they are.

Additionally, it is useful to conduct a sensitivity analysis regarding the chosen cut-level to investigate, whether the obtained TCs were the same at higher and lower cut-levels [26]. Therefore, in the present study the TT procedure was repeated at three cut-levels above and under the chosen one.

In a further step, individual component scores for all participants are calculated by summing the products of the original variable value and the variable loading on the treelet component in a linear way (FORMULA 4). These individual component scores can be used to associate TCs to different outcomes of interest.

$$\sum_{n=1}^{x} \text{value of standardised original variable [x] * treelet component load of original variable [x]}$$

FORMULA 4: Formula to calculate individual component scores for all participants; x: value of variable

In the present study, data on small polar metabolites (spmets) and lipids detected in serum samples were reduced by using the TT method. Due to non-normal distribution spmets variables were log transformed before analysis. Concerning lipid variables non transformed variables were normally distributed in more cases compared to log-transformed variables, therefore non-transformed lipid variables were used.

In both data sets of identified metabolites and lipids, respectively, variables with extremely high variances, i.e. the uppermost 10%, were excluded. For spmets it meant the exclusion of 13 variables (TABLE S 12) and for lipids 39 variables (TABLE S 13) were not included, resulting in 121 spmets and 353 lipids, respectively, which were used in the TT analysis.

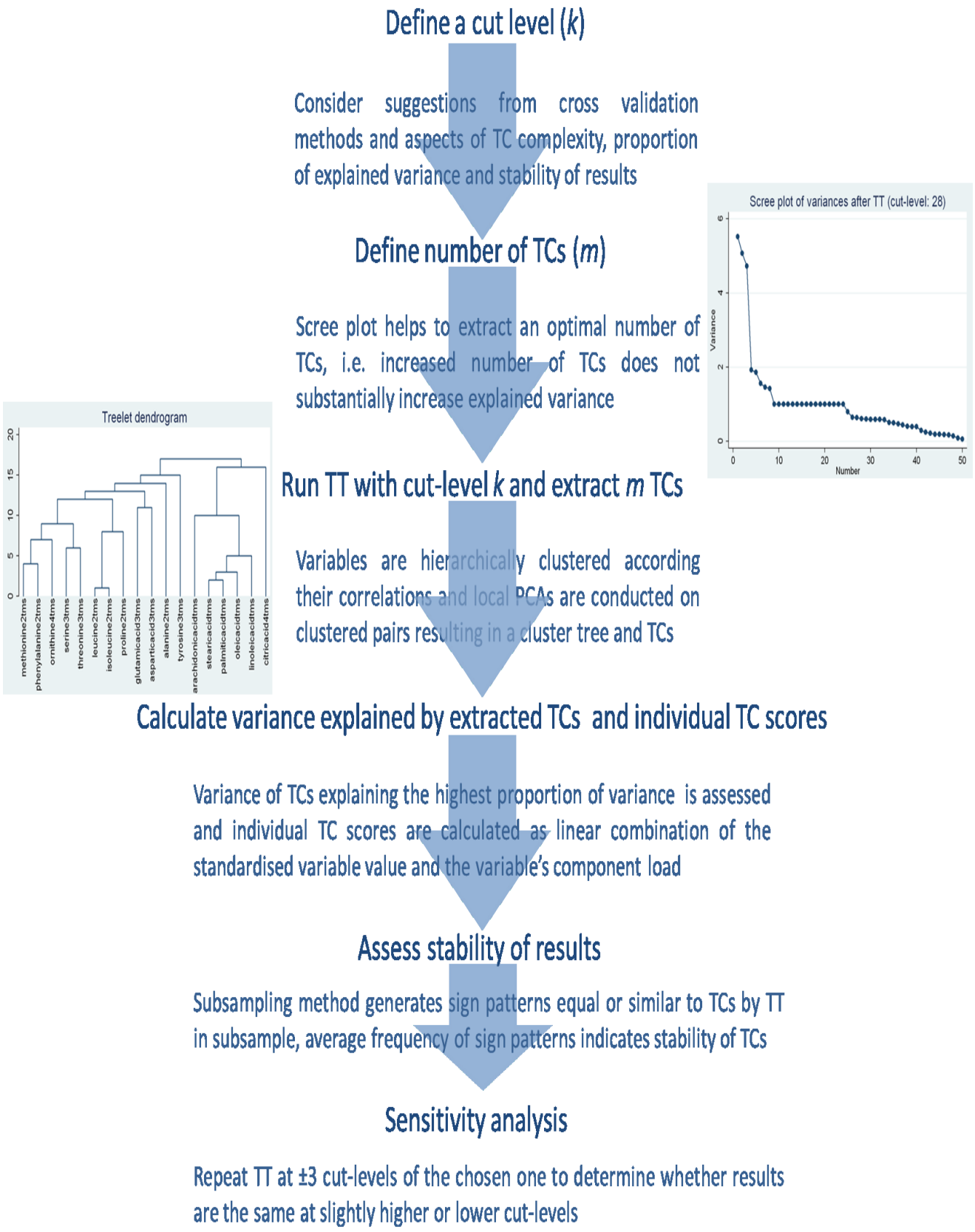


FIGURE 8: Flow chart indicating the general strategy of a TT procedure; TC-treelet component, TT-treelet transform

2.3.3 Dimension reduction with Principal Component Analyses (PCA)

Dimensionality of data on serum metabolites and serum lipids was also reduced through a PCA. PCA is an established method with the aim to reduce dimension in inter-correlated data by maximising the explained variance in all original variables [64]. It summarises the variables according to their correlations among each other and generates by this a smaller number of latent factors – the principal components, in the following consistently referred to factors. Each of the obtained factors is a linear combination of all original variables. Meaning that every original variable loads to a specific magnitude on all extracted factors, making the factors complex constructs. PCA generates as many factors as original variables are available. To reduce dimension it is necessary to decide for a suitable number of factors to be retained. There are several methods to consider the aspect of how many factors to retain [65]. In this thesis the scree plot criterion combined with aspects of non-trivial factors and percent of cumulative variance was used. In a scree plot the eigenvalues of all factors, which are proportional to the percentage of explained variance, are plotted depending on the number of factors. For a decision on the number of factors to extract, a 'big gap' in the scree plot has to be identified and thus, the number of components lying before this gap [66] is helpful. Adding factors beyond this 'gap' will not substantially increase the percentage of explained variance (FIGURE 9).

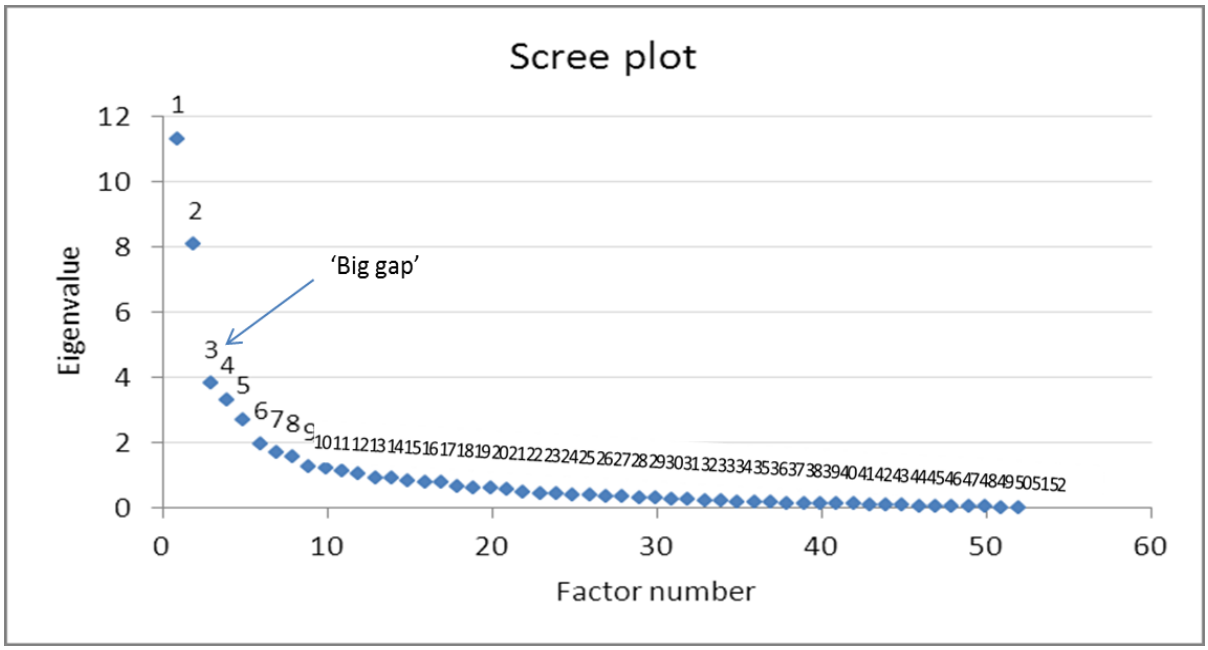


FIGURE 9: Scree plot from a fictional PCA on 52 original variables; between the second and the third factor a 'big gap' can be identified

To receive a simple structure and an improved interpretability of the results, factors where rotated using the *varimax* method as an orthogonal rotation procedure. There are several rotation methods being either orthogonal or oblique. Assuming that the obtained factors are uncorrelated, it was decided to run the orthogonal *varimax* procedure [67, 68] resulting in still orthogonal factors with a clearer structure, i.e. variables load either near 1 or near 0 on the factors. Variables that load high on the factor (near 1) are important and variables that load low (close to 0) are less important for the interpretation of the factor [67]. In the present study a variable loading higher than |0.35| on a factor was considered as an important variable. By determining the number of original variables that load on at least two of the factors the obvious interpretability of the extracted factors was evaluated.

The stability of the retained factors was assessed by a repetition of the PCA with the same characteristics in bootstrap-samples with the size of 80% of the original data set ($n_{bootstrap\text{-}samples}$ = 181). The extracted factors were compared with the original factors regarding their loadings. Stability was described by the percentage of how often an original variable loading on the initial factor appeared among the factors from the bootstrap-sample PCAs with a similar loading (±25%). Further it was assessed how many different variables load on the factors in the bootstrap-sample PCA with a loading higher than |0.35|.

To assess relations between obtained factors and outcomes of interest individual factor scores were computed for each participant. The SAS procedure PROC FACTOR was used for the PCA. Factor scores were calculated as the linear combination of the products of the z-standardised variable value (with mean = 0 and standard deviation = 1) and the standardised scoring coefficients [69]. The standardised scoring coefficients were computed by multiplying the inverse of the variable correlation matrix by the structure matrix.

2.3.4 Method comparison

To evaluate the applicability of the innovative dimension reduction method TT on the present metabolite data set, the findings were compared with the results of the PCA as an established dimension reduction method. The generated TCs were compared with the factors considering aspects of structure and percentage of explained variance. The structure was compared by using all non-zero loading variables of the TT and all high-loading (≥|0.35|) variables of the PCA. Explained variance of individual factors and cumulative percentage of explained variance were studied. Additionally, the methods were compared regarding the interpretability and the stability of retained TCs and factors in bootstrap-sample analysis.

2.3.5 Associations of treelet components and PCA factors with anthropometry and microbiota

Finally, the obtained TCs and factors were associated to anthropometric parameters to assess relations between serum metabolite and serum lipid profiles, respectively, and measures of obesity. Therefor, partial correlations [70] were calculated between the TCs or factors and body height, body mass, waist circumference, hip circumference, BMI, WHR, body fat mass, WHtR, BAI and body fat parameters from MRT analysis controlling for relevant covariates, such as age and sex [71, 72]. Anthropometric measurements, such as weight, hip circumference and BMI, existed for all participants thus 226 correlations could be calculated. MRT data were available from 177 participants (86 men and 91 women), therefore the correlations regarding visceral and subcutaneous fat could be only calculated for this group.

Also the association between the TCs or factors and the composition of the intestinal microbiota was investigated. As microbiota variable the dichotomous information on DGGE bands (0 vs. 1) was used. Regarding the TCs and factors the participants were divided into two groups depending on if their TC and factor score value was below or above the mean and the median value (1 vs. 2). Both variables are of dichotomous character therefore a φ-coefficient [73] was calculated to describe the association. Associations between microbiota bands and the mean and median groups, respectively, were calculated corrected for sex and age, since both covariates are relevant regarding these characteristics [71, 72, 74-76]. In the case that the association between the microbiota band and the mean group as well as the median group emerged to be significant the correlation between the microbiota and the TC or factor, respectively, was assumed to be relevant. Data on intestinal microbiota composition are existent for 216 participants (99 men 117 women), for which the correlations could be calculated.

Presented were the correlation coefficients between the microbiota bands and the mean groups of the TC or factor scores, respectively.

According to statements of J. Cohen [77], all significant correlations with a value $\geq|0.1|$ were considered as relevant regarding the relation of TCs or factors and anthropometry or intestinal microbiota.

Concerning the considerable number of missing values in the MRT and the microbiota data it was investigated, whether there is a disparity between the participants with and without the data in question. This sensitivity analysis contained investigation of age, sex, anthropometric parameters and TC scores.

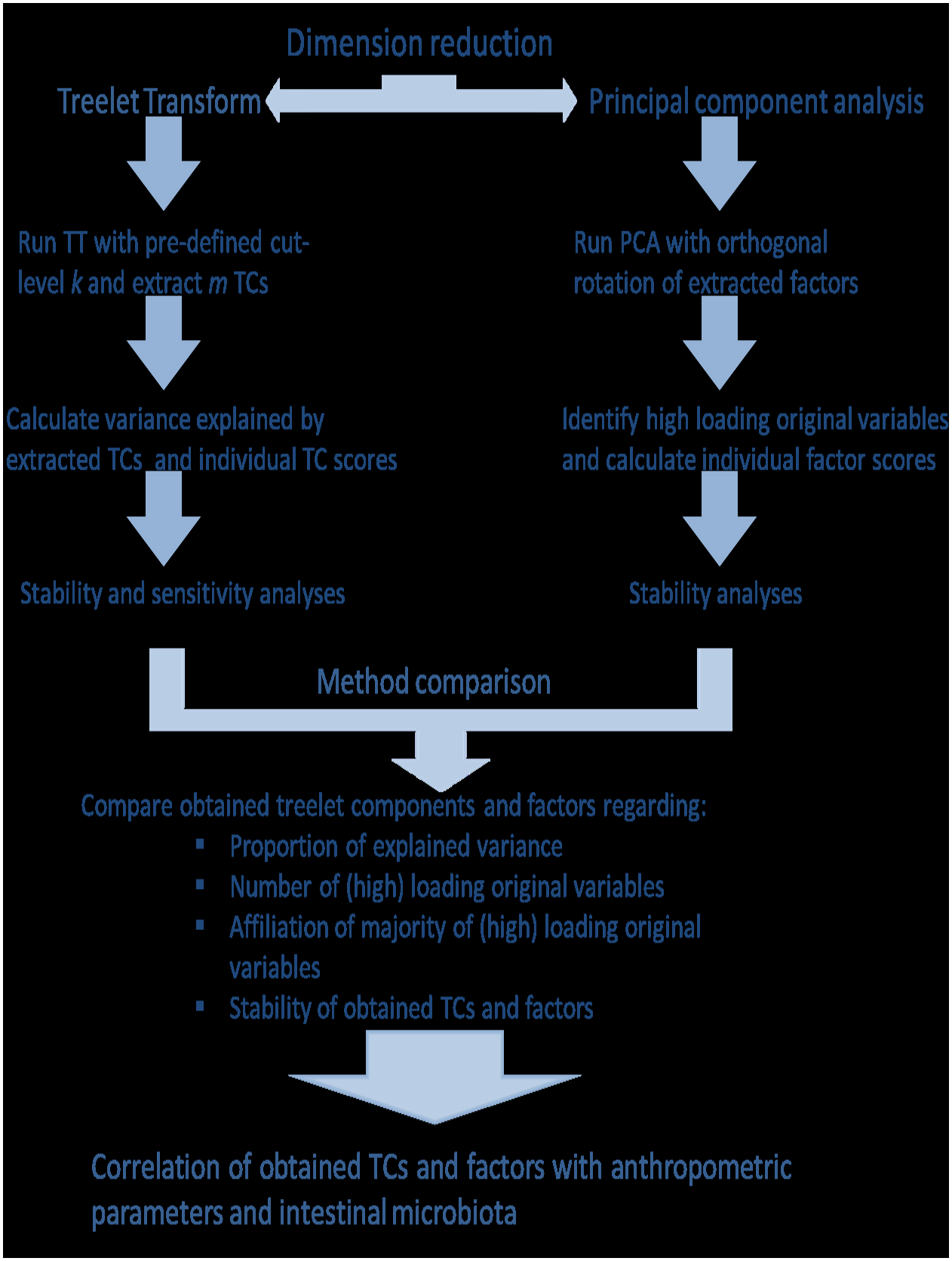


FIGURE 10: Flow chart indicating the overall strategy in present investigation

3. Results

3.1 Study population characteristics

The study population consisted of 120 female and 106 male elderly (mean age 63.6 years) participants (TABLE 2). Women were about 5 years younger than men due to a younger age at recruitment. On average participants could be categorised as overweight since they had a mean BMI of 27.3 kg/m². This was confirmed by the fact that only 15% of male and 19% of female participants could be classified in the lowest waist circumference category (below 80cm for women and below 94cm for men) being not associated to an increased risk for chronic diseases according to the WHO.

TABLE 2: Characteristics of the study population, stated are means and standard deviations (m±std); *n=177, 86 men 91 women; #n=216, 99 men and 117 women

	All (n=226)	Men (n=106)	Women (n=120)
Age [years]	63.6 ± 9.0	66.3 ± 8.6	61.2 ± 8.7
Body mass [kg]	77.4 ± 14.0	82.9 ± 12.9	72.5 ± 13.1
Body height [cm]	168 ± 9	174 ± 8	162 ± 6
Waist circumference [cm]	94.8 ± 12.6	99.7 ± 11.2	90.5 ± 12.3
Waist 1 ≤ 80/94cm	17	15	19
circumference 2 > 80/94 and ≤ 88/102cm	46	45	47
category [%] 3 > 88/102cm	37	40	34
Hip circumference [cm]	103.5 ± 9.2	101.4 ± 7.3	105.3 ± 10.3
Body fat mass [%]	31.2 ± 6.7	25.8 ± 4.4	35.9 ± 4.5
Visceral fat [cm³]*	4183.3 ± 2113.0	5236.1 ± 2108.0	3188.3 ± 1573.0
Subcutaneous fat [cm³]*	18834.5 ± 7218.8	15316.4 ± 5280.6	22159.4 ± 7250.2
Abdominal total fat [cm³]*	12058.9 ± 4909.4	11498.2 ± 4461.1	12588.9 ± 5268.0
visceral/abdominal subcutaneous fat ratio*	0.596 ± 0.327	0.866 ± 0.246	0.341 ± 0.127
BMI [kg/m²]	27.3 ± 4.2	27.3 ± 3.8	27.4 ± 4.6
WHR	0.92 ± 0.09	0.98 ± 0.06	0.86 ± 0.06
Waist-to-Height ratio	0.56 ± 0.07	0.57 ± 0.07	0.56 ± 0.08
Never smoker [%]	38	29	46
Active or moderately active [%]	56	56	54
Microbial diversity			
[nb of DGGE bands]#	27.7 ± 4.57	27.3 ± 4.67	28.1 ± 4.48
Shannon-Wiener index #	3.31 ± 0.18	3.29 ± 0.19	3.32 ± 0.17

Men and women had approximately the same amount of total abdominal fat with about 11500cm³ in men and 12500cm³ in women. Highest disparities could be observed in visceral fat, where the amount in women is lower (5236cm³ in men versus 3188cm³) whereas in subcutaneous fat women had much higher values (15316cm³ in men versus 22159cm³) (TABLE 2).

Female participants showed a slightly higher diversity in the intestinal microbiota with 28 different DGGE bands in the samples compared to the male participants with 27 different DGGE bands. Also regarding the Shannon-Wiener diversity index women had marginally higher values than men, 3.32 and 3.29, respectively. Overall 129 different bands could be detected in all samples indicating a high inter-individual diversity.

With 46% a higher proportion of women stated to be never smokers compared to men. Slightly more than 50% of all participants rated their scale of physical activity as active or moderately active.

3.2 Serum metabolites and lipids

The UPLC and GC x GC-TOFMS methods detected 1039 lipids and 587 small polar metabolites including 392 and 134 identified compounds, respectively. The mean, minimal and maximal value and variance of all compounds are summarised in TABLE S 1 to TABLE S 11.

Among the spmets a high proportion of the compounds could only be detected in parts of the study population. For 57% of the identified of the spmets (77 out of 134) the minimal value was 0.00, so these compounds were not detectable in at least one study participant. Most of the metabolites that were not detectable in all participants were fatty acids and carboxylic acid compounds (TABLE S 2 and TABLE S 3). Considering the serum lipid data only a few- in most cases polyunsaturated TGs (TABLE S 8) - were not detectable in the serum samples. Except for six lipids all others could at least be detected with a concentration of at least 0.01µmol/l in serum samples of all participants.

3.3 *Treelet transform* on identified small polar metabolites

First, the identified small polar metabolites were analysed with TT using the log-transformed variable. After exclusion of variables with extremely high variances (uppermost 10%, TABLE S 12), 121 metabolites were analysed. Cross-validation methods were performed to find the optimal cut-level for the obtained tree. Tenfold cross-validation with five and ten repetitions, respectively, was done for three to six components since this is a reasonable number of components to extract. All calculations yielded a cut-level between 81 and 87 (TABLE 3). A cut-level of 81 occurred twice in the cross-validation with ten repetitions and also for a number of five extracted components.

TABLE 3: Results of 10fold cross-validations to determine the optimal cut-level for dimension reduction with TT; cross-validations were done for three to six TCs

Nb of TCs	Cut-level 1	Cut-level2	Cut-level 3	Cut-level 4
		(5 repetitions)		(10 repetitions)
3	85	84	85	86
4	84	85	83	84
5	84	83	84	81
6	83	84	87	81

According to these results a TT on the correlation matrix of the 121 metabolites variables was performed with a cut-level of 81 and 5 extracted TCs. The cumulative explained variance of these TCs was 28%.

At this cut-level the TCs were very complex since they are loaded by many original variables. Also the results were not stable. Only two of the five extracted TCs occurred in more than 10% of the bootstrap-samples from the stability analysis. To find a more optimal cut-level, aspects of complexity and stability of TCs and proportion of explained variance were considered additionally. This resulted in a cut-level of 40 when extracting five TCs explained a total of 19% of variance within the metabolite data. To retain five TCs was decided according to the before mentioned criteria with focus on the scree plot (FIGURE 11). Extracting six TCs did not substantially increase the proportion of explained variance.

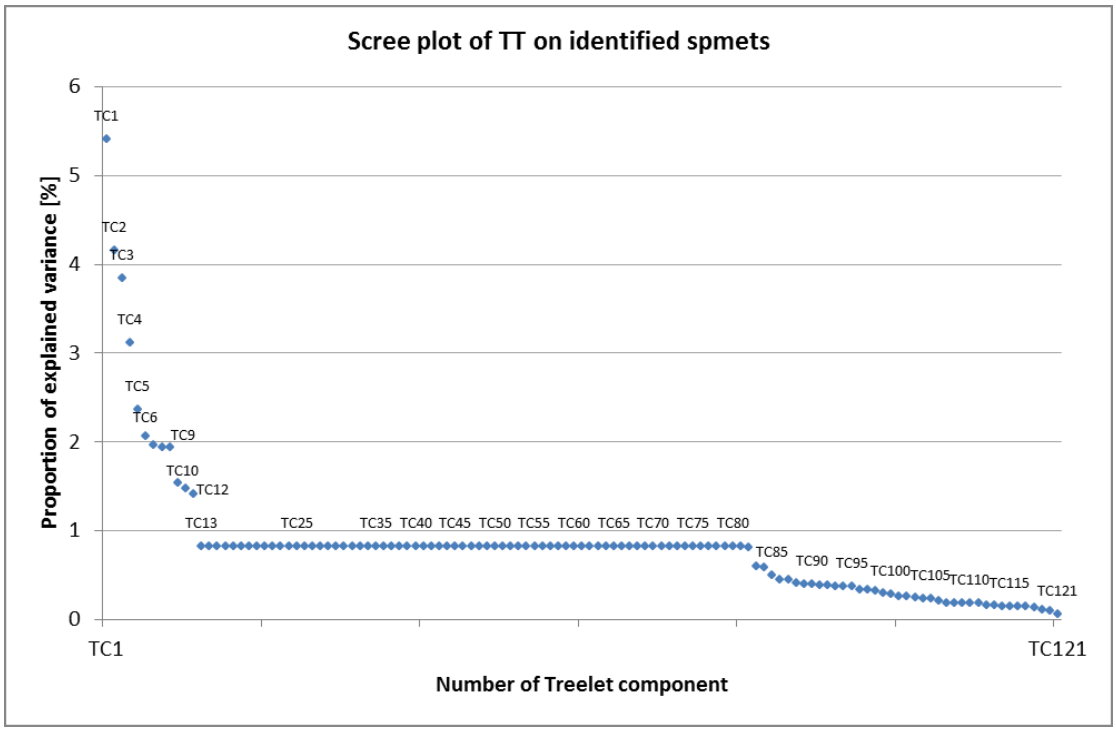


FIGURE 11: Scree plot of TT on identified small polar metabolites (p=121), retaining six TC_ms did not substantially increase proportion of explained variance, thus five TC_ms are retained

The single proportions of explained variance were 5.4, 4.2, 3.8, 3.1 and 2.4% for TC_m1 to TC_m5 with TC_m1 explaining the highest proportion (TABLE 4). TC_m1 was loaded by ten original variables of which six were amino acids, therefore TC_m1 was called amino acid component. TC_m2 was the fatty acid component since it was loaded by six fatty acids (out of seven original variables that loaded). TC_m3 contained eight sugar compounds and carboxylic acids and TC_m4 was loaded by five original variables of which four were amino acid derivates. Finally, TC_m5 comprised four original variables with three being sugar alcohols.

In all TC_ms original variables had positive loads. In addition, all results are illustrated in TABLE 4 and FIGURE 12 showing the obtained tree and the extracted TC_ms on the chosen cut-level.

TABLE 4: Characteristics and loading patterns of five extracted TCs generated by TT on 121 serum metabolite data

	TC_m1	TC_m2	TC_m3	TC_m4	TC_m5
Explained variance	5.4%	4.2%	3.8%	3.1%	2.4%
Loaded by (Number of original variables)	10	7	8	5	4
Name	Amino acids	Fatty acids	Sugar compounds, Carboxylic acids	Amino acid derivates	Sugar alcohols
	Loading patterns				
Metabolite name (variable ID nb)	**TC_m1**	**TC_m2**	**TC_m3**	**TC_m4**	**TC_m5**
4-Methyl-2-oxovaleric acid (192)	0.2587				
Glucopyranose (206)	0.2907				
Glutamic acid (2)	0.2747				
Glutamine (361)	0.3516				
Hypoxantine (390)	0.3012				
L-Tryptophan (256)	0.3322				
Methionine (1)	0.3226				
Ornithine (10)	0.3516				
Phenylalanine (14)	0.3226				
Tyrosine (13)	0.3416				
9-Tetradecenoic acid (360)		0.336			
Linoleic acid (7)		0.4003			
Oleic acid (12)		0.4015			
Palmitic acid (8)		0.4138			
Stearic acid* (270)		0.332			
Stearic acid* (530)		0.336			
Stearic acid (6)		0.4138			
Arabinofuranose (144)			0.3399		
Erythrose (480)			0.3597		
Hexadecanoic acid, 2,3-bishydroxy (305)			0.3515		
Nonanoic acid (89)			0.3601		
Octanoic acid (263)			0.3601		
Ribonic acid (118)			0.35997		
Stearic acid* (112)			0.3515		
Stearic acid* (163)			0.34453		
Alanine, phenyl, dl (103)				0.468	
Pyroglutamic acid (98)				0.4384	
l-Threonine (184)				0.3951	
Tryptophan (287)				0.4623	
Tyrosine* (64)				0.468	
3,8-Dioxa-2,9-disiladecane (123)					0.512
Myo-Inositol (32)					0.4968
Pentitol, 1-desoxy (241)					0.4785
Xylitol* (135)					0.512

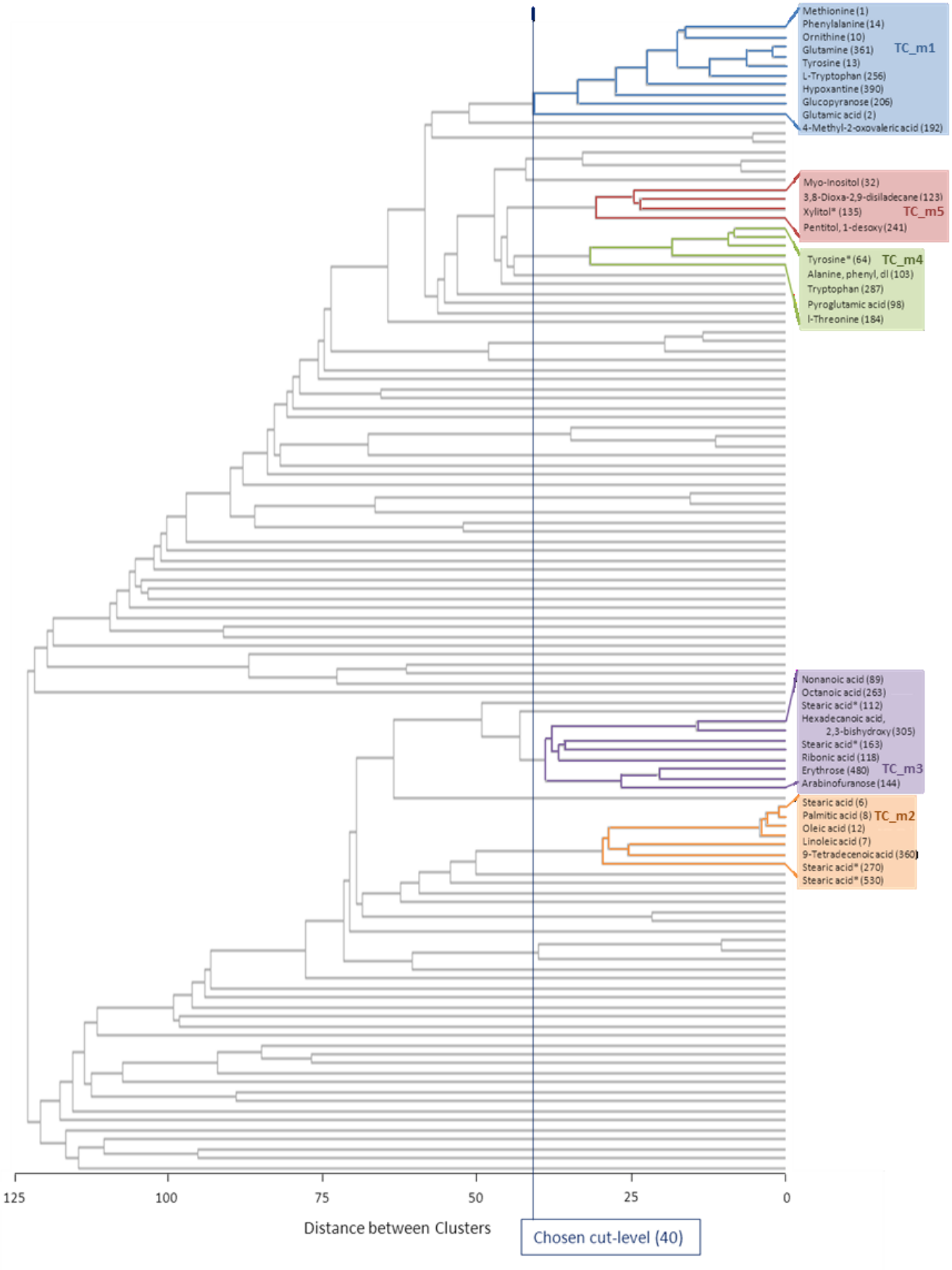


FIGURE 12: Dendrogram resulting from the TT analysis on 121 identified small polar metabolites, showing the chosen cut-level and the extracted TC_ms

Consistency of results was assessed in 80% bootstrap-samples with 100 replications in two stability runs. The first analysis resulted in eleven sign patterns that occurred in more than 10% of all bootstrap-samples. A sign pattern was a retained component with loadings similar or equal to the loadings of an original extracted TC. Sign pattern 2 corresponded exactly to TC_m1 whereas sign patterns 1 and 3 were similar to TC_m1. Together the three sign patterns were present in 75% of the bootstrap-samples (TABLE 5).

TABLE 5: Results of stability analyses of TT on identified small polar metabolites, bold are the sign patterns that match exactly to the TC_ms and their frequencies

TC_m	Similar sign patterns (changes compared to original TC_m)	Frequency
	First stability analysis	
1	1 (+1)[1], **2**, 3 (-1)[2]	12 + **36** + 27 → 75%
2	**4**	**78**%
3	**5**, 7 (-3)[3]	**36** + 11 → 47%
4	**6**	**71**%
5	8 (+1)[4], **10**	14 + **48** → 62%
	Second stability analysis	
1	**1**, 2 (-1)[2]	**42** + 34 → 76%
2	**3**	**76%**
3	**4**	**43%**
4	**5**	**80%**
5	**8**, 9 (+1)[4]	**51** + 13 → 64%

additional or missing variables in similar sign patterns [1]- Alpha-Ketoglutaric acid; [2]- 4-Methyl-2 oxovaleric acid; [3]- Ribonic acid, Erythrose, Arabinofuranose (all sugar compounds); [4]- Butane, 1,2,3-trishydroxy

TC_m2 was found in sign pattern 4 which accounted for 78% of all determined patterns. TC_m3 matched to sign pattern 5 and resembled sign pattern 7 with a frequency of 47%. The difference between TC_m3 and sign pattern 7 was that the sugar compounds were missing in the sign pattern. With a frequency of 71% sign pattern 6 described TC_m4. TC_m5 was equal to sign pattern 10 and similar to sign pattern 8 which together were present in 62% of the bootstrap-samples. Sign patterns 9 and 11 showed no similarities to any of the extracted TC_ms. The second stability analysis generated equal findings with nine sign patterns occurring in more than 10% of the bootstrap-samples. The lower number of sign patterns resulted from the circumstance that TC_m1 was described only by sign pattern 1 and 2, that had a frequency of together 76%, and that TC_m3 equalled only one sign pattern (sign pattern 4) being present in 43% of the bootstrap-samples (TABLE 5). In this stability analysis again two sign patterns (6 and 7) were not similar to any of the obtained TCs. The

frequencies of the resembling or matching sign patterns were similar to that from the first stability analysis. The analyses showed that TC_m1, m2 and m4 are extracted very reliably with relative frequencies of 71 to 80% in the subsamples, whereas TC_m3 seems to be relatively unstable occurring in only 43 and 47% of the subsamples, respectively.

Additionally, the choice of cut-level was verified in some sensitivity analyses by conducting a TT with all the original settings but with changing the cut-level by ±1 to ±3, so with the cut-levels 37, 38, 39, 41, 42 and 43. Using these cut-levels again 5 TC_ms were extracted explaining between 17.4% and 19.5% of the variance within the data (TABLE 6).

TABLE 6: Results of sensitivity analyses of the TT on identified small polar metabolites with cut-levels at 37, 38, 39, 41, 42 and 43

Cut-level	Explained variance [%]	Number of loading original variables				
		TC_m1	TC_m2	TC_m3	TC_m4	TC_m5
40 (chosen)	18.89	10	7	8	5	4
37	17.43	9[1]	7	5 (TC_m4)	5 (TC_m3)[2]	4
38	18.56	9[1]	7	8	5	4
39	18.56	9[1]	7	8	5	4
41	18.89	10	7	8	5	4
42	19.21	10	9 (TC_m3)[3]	7 (TC_m2)	5	4
43	19.54	10	9 (TC_m3)[3]	7 (TC_m2)	6[4]	4

Additional or missing variables in TCs at different cut-levels: [1]- 4-Methyl-2 oxovaleric acid; [2]- Ribonic acid, Erythrose, Arabinofuranose (all sugar compounds); [3]- Glyeric acid; [4]- Ethanolamine

The results from the TT with a cut-level of 41 were exactly the same compared to the original analysis with a cut-level at 40. At all cut-levels extracted TC_m2 and TC_m5 did not change. At levels below 40 in TC_m1 the 4-Methyl-2 oxovaleric acid was missing, but with levels above 40 TC_m1 was as the original one. In TC_m3 three compounds were missing at a level of 37, whereas at the levels 42 and 43 glyceric acid was additionally loading on TC_m3. At the cut-levels in between, TC_m3 was unchanged. The only cut-level at which

TC_m4 did not match the original TC_m was at a level of 43; here TC_m4 was loaded by six instead of five metabolites (TABLE 6).

One issue that was different on cut-level 37, 42 and 43 was the variance explained by the single TC_ms and by this the order of the TC_ms. At cut-level 37 TC_m3 and TC_m4 changed the succession; here TC_m4 became the TC explaining the third most variance. At levels 42 and 43 this was the case for TC_m2 and TC_m3; at both levels TC_m3 explained more variance than TC_m2 and ascended by this to the second place of TCs.

These sensitivity analyses confirmed the findings from the stability analyses with TC_m3 being the most unstable component with the sugar compounds being the changing part and with TC_m1, _m2 and _m4 being reliable components. TC_m5 remained stable in these analyses due to its large distance to the parental node in the tree (FIGURE 15).

3.4 Principal component analysis on identified small polar metabolites

To compare the TT results with an established dimension reduction method, a PCA on the same data set of log-transformed metabolite data (p=121) was conducted. The scree plot (FIGURE 13) indicated that extracting five instead of four factors would not substantially increase explained variance, therefore the extraction of four factors was chosen. In TABLE 7 the characteristics and loading patterns of these four factors are shown.

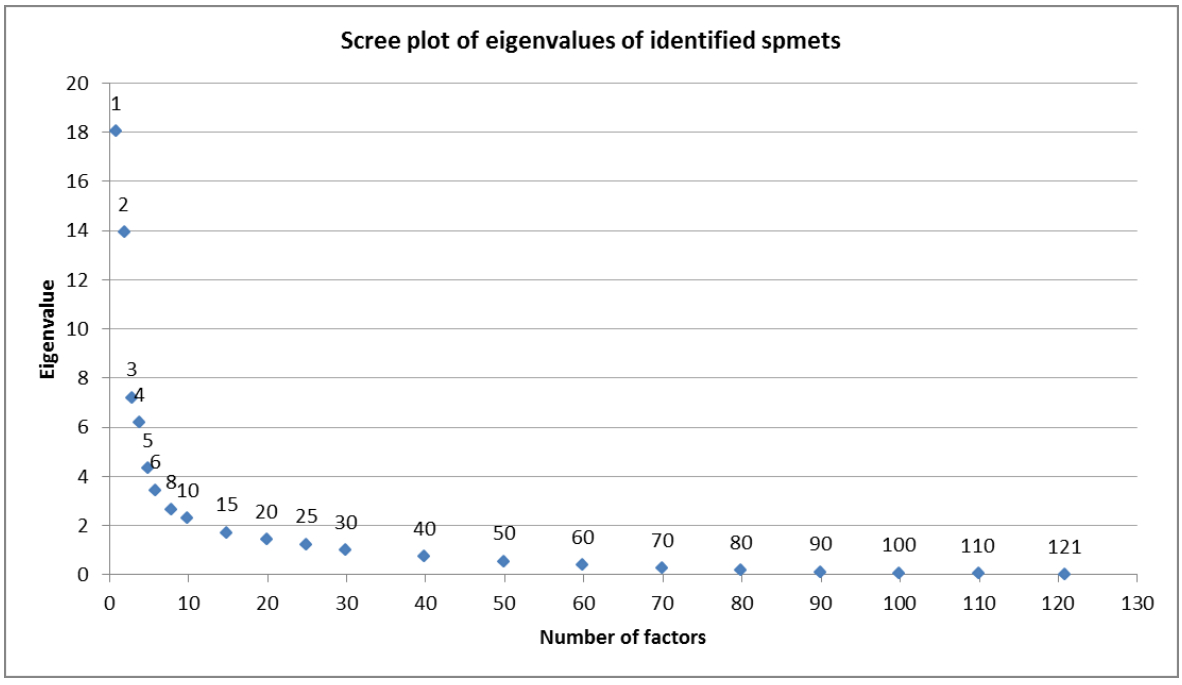


FIGURE 13: Scree plot of eigenvalues from PCA on identified spmets (p=121)

TABLE 7: Characteristics and loading patterns of four factors generated by a PCA on 121 serum metabolite variables

	Factor_m1	Factor_m2	Factor_m3	Factor_m4
Explained variance	14.9%	11.5%	6.0%	5.1%
Loaded higher than |0.35| by (number of original variables)	41	30	24	24
Name	Amino acid derivates	Sugar compounds, Carboxylic acids	Fatty acids	Amino acids
Number of loading variables that load in at least one more factor	15	8	11	12
		loading patterns		
Metabolite name (ID nb)	**Factor_m1**	**Factor_m2**	**Factor_m3**	**Factor_m4**
Myo-Inositol (32)	0.78			
Alanine (103)	0.75			
L-Proline (98)	0.74			
Xylitol (135)	0.71	0.4	0.55	
Butanoic acid (78)	0.71			
Tyrosine (64)	0.7			
2-Butenedioic acid (107)	0.69			
Malic acid (95)	0.67			
3,8-Dioxa-2,9-disiladecane (123)	0.63			0.47
Tryptophan (287)	0.61		0.51	
Xylitol (102)	0.59			0.45
Lactic acid (26)	0.59			
Adipic acid (288)	0.58	-0.66		
Silanamine (79)	0.58	-0.7		
Stearic acid (112)	0.56			
Glyceric acid (85)	0.56			
Butane (1187)	0.54	0.62		
Trimethylsilyl ether of glycerol (149)	0.54		0.41	
(R*,R*)-2,3-Dihydroxybutanoic acid	0.53			
Pentitol (241)	0.53	-0.54		
?-Alanine (175)	0.52			-0.48
l-Threonine (184)	0.52		0.42	
Decanoic acid (33)	0.51	0.62	0.38	
2-Hydroxybutyric acid (294)	0.49			
d-Galactose (337)	0.49			
4-Trimethylsiloxyvalerate (590)	0.45			
Trimethylsiloxyproline (88)	0.45			
Phosphoric acid (264)	0.44			
Octanoic acid (263)	0.43			
4-Methyl-2-oxovaleric acid (192)	0.41			
Hexadecanoic acid (305)	0.4			
L-Proline (137)	0.39			
2-Ethyl-3-hydroxypropionic acid (284)	0.37			
3-Trimethylsiloxycaproate (409)	0.37			
2,3,4-Trihydroxybutyric acid (47)	0.37			
2,5-Furandicarboxylic acid (556)	0.37			
Nonanoic acid (89)	0.37	0.42		
2-Furancarboxylic acid (600)	0.36			0.37

1H-Indole-3-acetic acid (63)	0.36	0.59		
Glucopyranose (206)	0.35			
D-Glucitol (417)	0.35			
Ribonic acid (118)		0.88		0.54
Erythrose (480)		0.8	0.83	
Arabinofuranose (144)		0.74	0.77	
Arachidonic acid (3)		0.64		
Heptanoic acid (438)		0.57		
Stearic acid (6)		0.44		
Stearic acid (163)		0.43		0.64
Arachidonic acid (1346)		0.42		
Ribonic acid (1189)		0.39		
Linoleic acid (134)		0.38		0.38
Palmitic acid (8)		0.36		0.45
Linoleic acid (157)		0.35		
Propanoic acid (282)		-0.37	0.42	
Glutamic acid (2)		-0.41	0.37	0.43
Phenylalanine (14)		-0.41		
Glucopyranose (167)		-0.42		0.64
Tyrosine (13)		-0.55	0.45	
Methionine (1)		-0.56	0.63	
Glutamine (361)		-0.6		0.56
Ornithine (10)		-0.61		
Alpha-ketoglutaric acid (540)		-0.67		
L-Tryptophan (256)		-0.71		0.67
Oleic acid (12)			0.88	
Linoleic acid (7)			0.78	
9-Tetradecenoic acid (360)			0.78	
Stearic acid (530)			0.74	
Stearic acid (270)			0.7	
Linolenic acid (159)			0.6	
Linoleic acid (1185)			0.55	
3-Hydroxybutyric acid (21)			0.55	
Hexadecanoic acid (300)			0.55	
Lauric acid (179)			0.52	
Arachidonic acid (1192)			0.38	
Arachidonic acid (3)			0.38	
Alanine (20)			-0.38	
Hypoxantine (390)				0.7
Creatinine (54)				0.67
Serine (4)				0.59
1,3-Benzenedicarboxylic acid (397)				0.58
Leucine (15)				0.53
Isoleucine (17)				0.52
Proline (19)				0.52
Threonine (9)				0.52
Aspartic acid (18)				0.48
Valine (23)				0.48
Hippuric acid (462)				0.37
tert-Butylpentamethyldisiloxane (467)				-0.47

The proportions of explained variance were 14.9%, 11.5%, 6.0% and 5.1% for the factors factor_m1 to factor_m4 (TABLE 7) with respective eigenvalues of 18.1, 13.9, 7.2, and 6.2 and an overall proportion of explained variance of 37.5%. Variables loading higher than 0.35 or lower than -0.35 were considered as relevant for the interpretation of the factors. According to that factor_m1 was loaded by 41 variables, a large part of them being amino acid derivates. Factor_m2 was loaded by 30 original variables that were mostly sugar compounds and carboxylic acids, respectively. Factor_m3 and factor_m4 were both loaded by 24 variables. For factor_m3 these were fatty acids and amino acids for factor_m4 (TABLE 7).

50% (12 out of 24) of the high loading variables of factor_m4 also load high on other factors, comparable values could be determined for all other factors with 37%, 27% and 46%, respectively for factor_m1 to factor_m3 (TABLE 7). This depicts an issue that complicates the distinct interpretation of the factors.

PCA factors showed an intermediate stability in bootstrap-samples (TABLE 8), the original variables loading on the initial factors occurred on average in 54% of the factors from a PCA on bootstrap-samples with approximately the same loading. The most instable factor was factor_m2 with an average frequency of the initially loading variables in the bootstrap PCA factors of 37.5%. There even was one variable, Glycopyranose (167), that did not load in one of the bootstrap PCA within a range of 75% to 125% of the original loading. In all factors some variables occurred quite seldom in the bootstrap PCA factors. The highest values could be found for factor_m4 whereby the initially loading variables could be found in at least 40% of the bootstrap factors (TABLE 8).

TABLE 8: Results of stability analyses of the PCA regarding the identified small polar metabolites

Factor	Nb of loading original variables	Nb of loading original variables in bootstrap PCAs	frequency of original variables in bootstrap PCAs with loading in the same range
_m1; Amino acid derivates	41	84	58.8 % (10 - 90%)
_m2; Sugar compounds, Carboxylic acids	30	94	37.5% (0 - 60%)
_m3; Fatty acids	24	63	57.9% (10 - 90%)
_m4; Amino acids	24	63	60% (40 - 90%)

Overall the factors extracted from the PCAs in bootstrap-samples were loaded ≥|0.35| by much more original variables than the factors from the original PCA, e.g. factor_m1 originally loaded by 41 variables was loaded by 84 different variables in the bootstrap factors. Also for

all other factors (_m2, _m3 and _m4) the fact persisted that more than twice as many original variables loaded on the bootstrap factors compared to the initial PCA factor (TABLE 8).

3.5 Comparison of *treelet transform* and principal component analysis on identified small polar metabolites

A comparison of structures of the factors obtained from PCA and components from TT showed that all factor_ms could be found among TC_ms (TABLE 9).

Regarding the proportion of explained variance the factors retained from PCA explained nearly twice as much compared to the TCs from TT. This ratio could also be found when the extracted factors and TCs were considered separately. A factor from PCA always explained at least twice as much variance as a factor. In terms of the first TC and factor the difference was even higher. TC_m1 explained 5.4% (TABLE 4) and factor_m1 14.9% (TABLE 7), so almost three times as much.

TABLE 9: Characteristics of the TT and the PCA conducted with the same data set of small polar metabolites

	TT	**PCA**
overall explained variance	18.9%	37.5%
Number of extracted factors/components	5	4
Component/Factor (number of (high) loading variables)		
Amino acids	TC_m1 (10)	Factor_m4 (24)
Fatty acids	TC_m2 (7)	Factor_m3 (24)
Sugar compounds, carboxylic acids	TC_m3 (8)	Factor_m2 (30)
Amino acid derivates	TC_m4 (5)	Factor_m1 (41)
Sugar alcohols	TC_m5 (4)	not available
Stability		
Frequency in bootstrap investigation*	67.2%	42.8%
[mean, min – max]	(43% – 80%)	(38% - 60%)

* for TT frequency of similar sing patterns in TC from bootstrap-samples is determined; for PCA it is the frequency of initially loading variables in bootstrap PCA factors

TCs from TT were more stable compared to factors from PCA on spmets. Bootstrap investigations showed that 67.2% of the TC occurred with the same or similar sign patterns

among the bootstrap TCs. Original variables that load on initial PCA factors occurred with a similar loading in only 42.8% of the factors from bootstrap PCAs (TABLE 9).

Further, the difference was noticeable with regard to the number of loading original variables. In all cases the TC was loaded by less than half of the variables compared to the correspondent factor (TABLE 9). The most obvious was the difference in the amino acid derivate component and factor. While TC_m4 was loaded by five original variables, factor_m1 was loaded by 41 variables with a loading ≥|0.35|. The smallest difference could be observed for the amino acid component and factor. TC_m1 was loaded by ten variables and factor_m4 by 24 original variables.

Original variables that loaded on TC_m1 also were found in factor_m4 with a few exceptions. Both, TC_m1 and factor_m4, were dominated by amino acids (FIGURE 14). Since TC_m1 was loaded by ten original variables, out of them seven amino acids, factor_m4 consisted of 24 original variables loading higher than |0.35|, out of them 15 amino acids. The further loading variables contained amino acid derivates, fatty acids and other carboxylic acids.

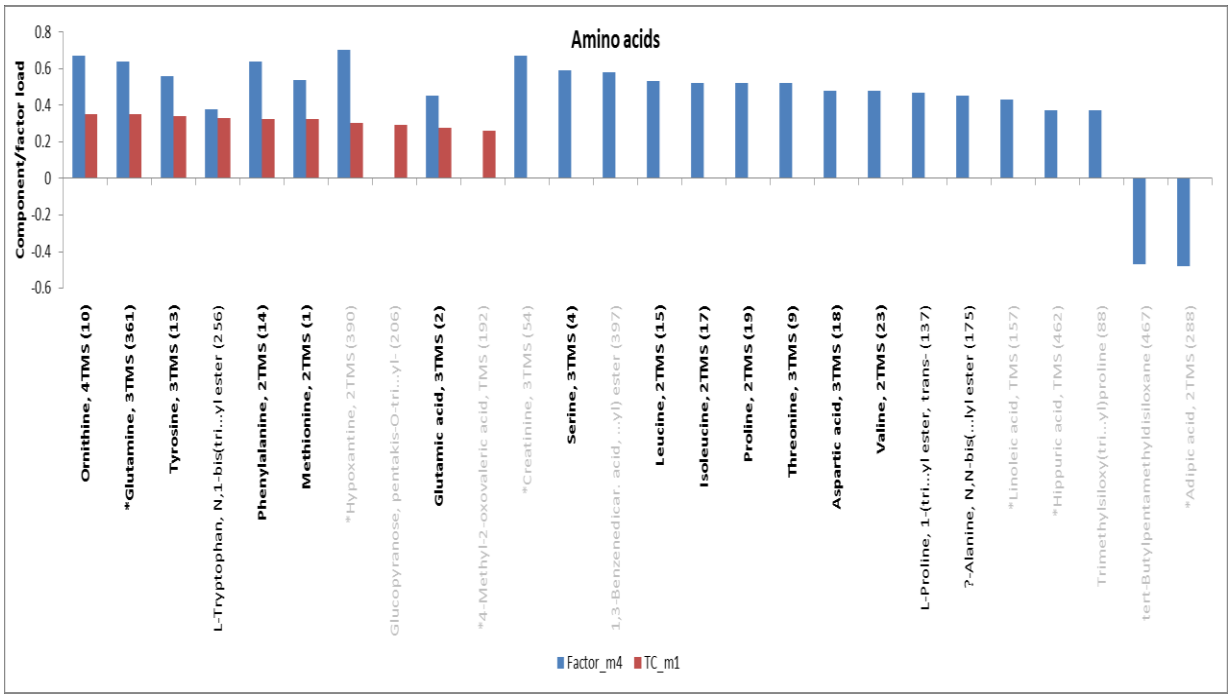


FIGURE 14: Comparison of the loading structure of TC_m1 and factor_m4, the amino acid component and factor; shown are all original variables loading on TC_m1 and on factor_m4 (higher than |0.35|), bold are the amino acids, normal the related compounds and grey all other compounds

Similar observations were made for TC_m2 and factor_m3, both were mainly loaded by fatty acids or derived compounds (FIGURE 15).

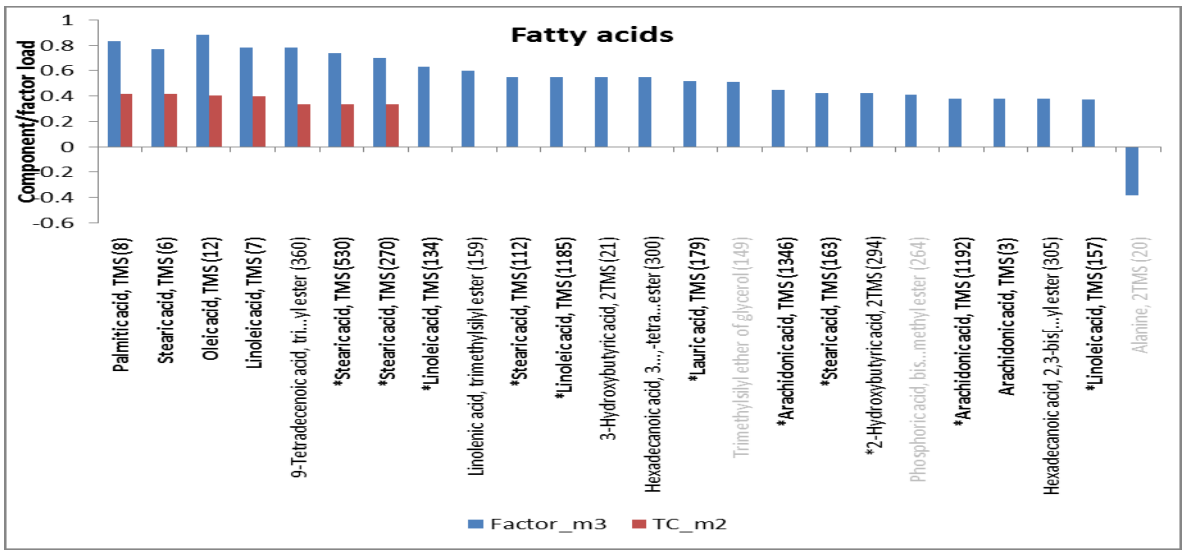


FIGURE 15: Comparison of the loading structure of TC_m2 and factor_m3, the fatty acid component and factor; shown are all original variables loading on TC_m2 and on factor_m3 (higher than |0.35|), bold are the fatty acids, normal the related compounds and grey all other compounds

The sugar compounds and carboxylic acids that loaded on TC_m3 also were present in factor_m2 (FIGURE 16).

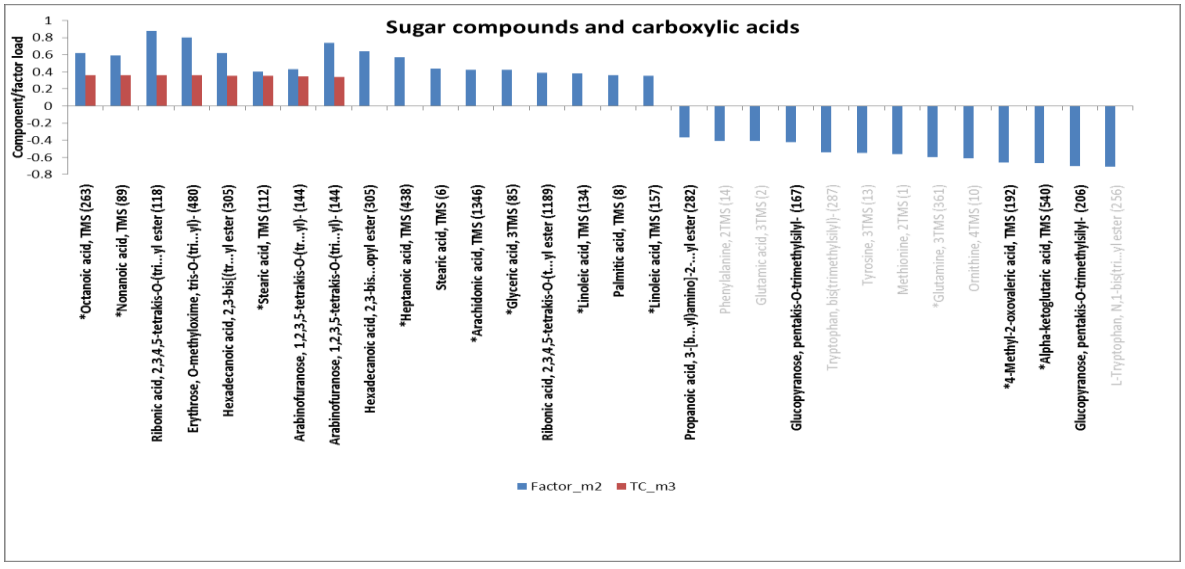


FIGURE 16: Comparison of the loading structure of TC_m3 and factor_m2, the sugar compound and carboxylic acid component and factor; shown are all original variables loading on TC_m3 and on factor_m2 (higher than |0.35|), bold are the sugar compounds and carboxylic acids, normal the related compounds and grey all other compounds

The most complex factor from the PCA was factor_m1 with 41 corresponding variables among them all amino acid derivates that also described and loaded on TC_m4 (FIGURE 17).

Further all compounds that loaded on TC_m5 (sugar alcohols) were present among the loading variables of factor_m1 (FIGURE 17).

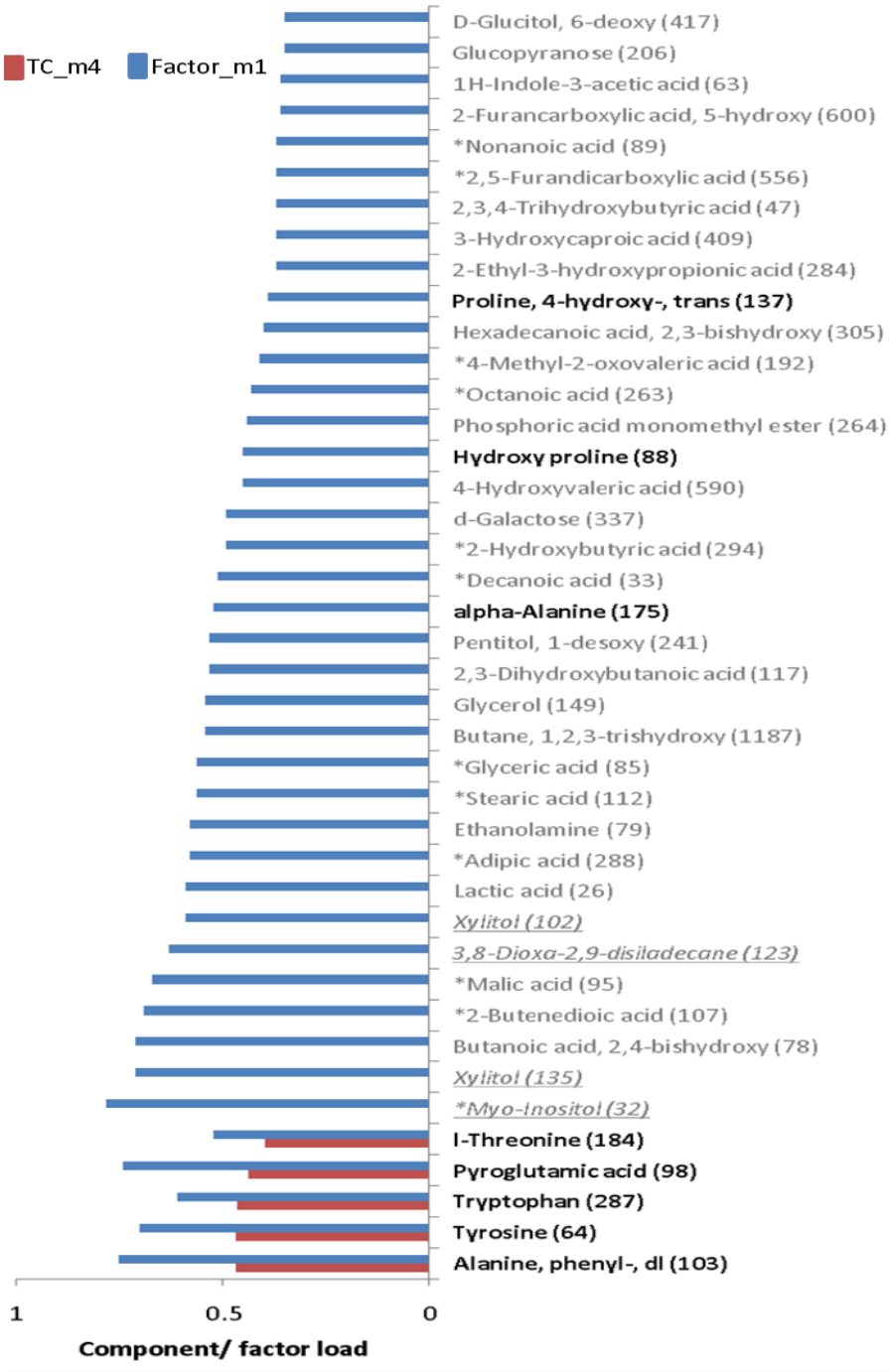


FIGURE 17: Comparison of the loading structure of TC_m4 and factor_m1, the amino acid derivate component and factor; shown are all original variables loading on TC_m4 and on factor_m1 (higher than |0.35|), bold are the amino acid derivates and grey all other compounds, additionally underlined and italic are the compounds loading on TC5 (sugar alcohols)

3.6 Associations of metabolite treelet components and factors with anthropometry

To assess the relationship between serum metabolites and anthropometric parameters, partial correlations between the obtained components and anthropometric measurements were calculated also considering age and sex as influencing factors. Both, TC_m1 and factor_m4, containing mainly amino acids, were positively correlated with almost all investigated anthropometric parameters, in particular with body weight, waist circumference, BMI, WHtR and body fat mass (TABLE 10). Especially the association with body fat mass was strongest with correlation coefficients of 0.15 and 0.22 for TC_m1 and factor_m4, respectively. The TC_m1 additionally was directly correlated with hip circumference, whereas the factor_m4 was further associated with WHR.

TABLE 10: Partial correlations (corrected for age and sex) between obtained TC_ms or PCA factor_ms and anthropometric parameters; BMI- body mass index, WHR- waist-to-hip ratio, WHtR- waist-to-height ratio, BAI-body adiposity index

	Treelet component/ PCA factor	Weight [kg]	Waist circumference [cm]	Hip circumference [cm]	BMI [kg/m²]	WHR	Body fat mass [%]	WHtR	BAI	visceral fat [cm³]	subcutaneous fat [cm³]	abdominal total fat [cm³]	viscerl/abdominal subcutaneous fat ratio
Amino acids	TC_m1	0.126	0.111	0.112	0.125		0.145	0.101		0.116	0.128	0.148	
	Factor_m4	0.104	0.121		0.112	0.116	0.224	0.119					
Fatty acids	TC_m2												
	Factor_m3												
Sugars and carboxylic acids	TC_m3												-0.110
	Factor_m2												
Amino acid derivates	TC_m4										0.137	0.119	
	Factor_m1				0.117					0.109	0.208	0.159	
Sugar alcohols	TC_m5								0.105		0.134	0.103	

Degree of correlation	<= -0.2	-0.2 to -0.13	-0.129 to - 0.1	0.1 to 0.129	0.13 to 0.2	>= 0.2

Also with some of the body fat data from MRT analysis associations of TC_m1 could be observed. The component was moderately correlated with subcutaneous, visceral and abdominal total fat mass (r= 0.13, r= 0.12 and r= 0.15, respectively) (TABLE 10). The corresponding factor_m4 was not correlated with any of the MRT data.

Further, infrequent correlations could be observed for the TC_m and factor_m described by amino acid derivates. TC_m4 and factor_m1 were both associated with subcutaneous and abdominal total fat and the factor was also directly related to visceral fat and BMI (TABLE 10).

Regarding the component and the factor loaded by sugars and carboxylic acids, only TC_m3 showed a correlation to the ratio from visceral and abdominal subcutaneous fat, here the correlation coefficient was -0.11.

Also the sugar alcohol TC, TC_m5; displayed only single associations to the anthropometric measurements. It was correlated to BAI (r=0.11), to subcutaneous fat (r=0.13) and to abdominal total fat (r=0.10).

The factor and component reflecting fatty acids (factor_m3 and TC_m2) or sugars and carboxylic acids (factor_m2 and TC_m3) did not show any relation to anthropometric measures, except slightly negative correlation between TC_m3 and the ratio of visceral and abdominal subcutaneous fat (r=-0.11).

Also body height was not correlated with any of the investigated serum metabolite components and factors.

3.7 Associations of metabolite treelet components and factors with intestinal microbiota

All of the extracted TC_ms and factor_ms were linked with some of the described DGGE microbiota bands. TABLE 11 shows the association between the factor_ms or the TC_ms, respectively, and the microbiota bands. The bands being associated with the factor_m as well as with the TC_m are accentuated in bold. For the factor and the TC characterised by amino acid derivates (TC_m4 and factor_m1) and by fatty acids (TC_m2 and factor_m3) this was the case for one band (b700 and b509, respectively).

TABLE 11: Correlations of metabolite factors and TCs with intestinal microbiota, stated are φ-coefficients for pairs with significant correlation ($p < 0.05$)

TC or factor \DGGE band	Amino acids		Fatty acids		Sugar compounds		Amino acid derivates		Sugar alcohols	
	TC_m1	Factor_m4	TC_m2	Factor_m3	TC_m3	Factor_m2	TC_m4	Factor_m1	TC_m5	
b188	0.134		b099	0.207	**b188**	-0.205	-0.179	b198	-0.191		b300	-0.168
b313	0.183		**b509**	0.135	0.192	b313	-0.221		b300	-0.139		b306	0.159
b329	-0.200		b513	-0.179		b329	0.153		b341	-0.144		b368	-0.155
b341	-0.186		b740	0.159		b432	0.149		b378	-0.153		b472	0.199
b388	-0.162		b953	0.145		**b610**	-0.180	-0.178	b388	-0.207		b523	-0.162
b625	0.208		b958	0.201		**b879**	-0.160	-0.213	b394	0.138		b879	0.199
b710	-0.149		b496		0.173	b191		-0.162	b509	-0.141			
b921	-0.205	-0.174	b625		-0.186	b450		0.149	b554	-0.163		1*	
b962	0.168		b700		-0.151	b533		-0.148	b571	-0.193		2**	
b417		-0.192	b757		0.207	b740		0.166	b690	-0.179			
diversity									**b700**	0.152	0.147		
1*			1*			1*			b725	-0.143			
2**			2**			2**			b901	0.171			
									b466		-0.146		
									b472		0.212		
									b523		-0.161		
									b583		-0.159		
									b640		-0.207		
									b745		-0.143		
									b817		0.156		
									1*				
									2**				
Degree of correlation	<-0.2	-0.2 to -0.13	-0.1 to -0.129	>0.2	0.13 to 0.2	0.1 to 0.129				

bold are the DGGE bands which are associated with the factors and the TCs; 1*- diversity as total number of DGGE bands per sample and 2**- Shannon-Wiener diversity index

For most common associations the φ- coefficient between the DGGE band and the TC_m was stronger than that with the factor_m, but in all cases the tendency of the correlation was the same. For the amino acid derivates component and factor and for the fatty acid component and factor the common associations were positive. Factor_m2 and TC_m3 (sugar compounds) showed three overlaps regarding the correlation with intestinal microbiota, the DGGE bands b188, b610 and b879 were correlated with the factor_m and with the TC_m.

Again, the tendency of the correlation was the same for the factor_m and the TC_m, all bands were negatively correlated. The amino acid factor and TC (factor_m4 and TC_m1) both had a negative correlation with one microbiota band (b921). Here the φ-coefficient for the association with b921 was again stronger for the TC_m1 (-0.205) compared to the association to factor_m4 (-0.174). The sugar alcohols component, which only could be extracted as TC_m, was correlated to six microbiota bands. Overall, the significant correlations had an extension between 0.134 and 0.221 with a mean of 0.172 for the absolute values. No significant correlation was observed between the metabolite factors or TCs and the microbial diversity, expressed as the total number of detected DGGE bands per sample or the Shannon-Wiener diversity index. All calculated correlation coefficients were in the range between -0.11 and 0.04 for both diversity parameters.

3.8 *Treelet transform* on identified serum lipids

For TT analysis on serum lipid data the uppermost 10% of variables with high variances were excluded (TABLE S 13). Thus, 353 lipid variables were analysed. Variables were not log transformed, since they are closer to normal distribution in the non-transformed state. The cross-validation resulted in an optimal cut-level of 325 for three retained components (Cross-validation results from *treelet transform* on serum lipids TABLE S 14). A TT calculated for three components and with a cut-level of 325 generated very complex and quite unstable TCs (TABLE 12). Up to 140 original variables load on the TCs obtained by a TT with the mentioned characteristics. Further, sign patterns similar to TC_l2 and TC_l3 occurred in less than 5% of the TCs resulting from stability analyses on bootstrap-samples and sign patterns similar to TC_l1 occurred in 22 and 23%, respectively, in the two stability runs (TABLE 12).

TABLE 12: Results of TT with a cut-level of 325 and three retained components on 353 serum lipid variables

	Explained variance	**Number of loading original variables**	**Occurrence of similar sign patterns in subsamples**	
			1st stability run	2nd stability run
TC_l1	15.9%	140	23%	22%
TC_l2	8.8%	84	< 5%	< 5%
TC_l3	4.4%	32	< 5%	< 5%

Thus a lower cut-level of 200 was chosen for further analyses. The cut-level took aspects of component complexity and by this interpretability, proportion of explained variance and stability of the results into consideration. Cutting the dendrogram generated by the TT algorithm at a cut-level of 200 produced three components (FIGURE 19 and TABLE 13) explaining 17.8% of the variance within the serum lipid data.

The TC_l1 was loaded by 43 original variables all of them triglycerides and 33 saturated or monounsaturated which is why that component was called saturated and monounsaturated TGs. TC_l2 was mostly loaded by polyunsaturated TGs (23 out of 29) and 77% of the 13 variables loading on TC_l3 were PCs and PEs (TABLE 13).

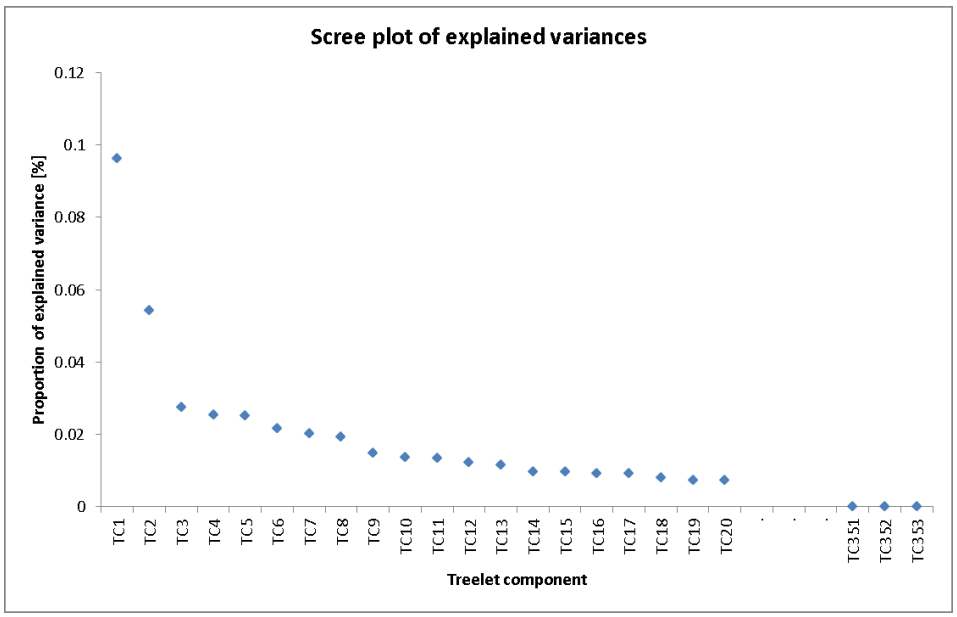
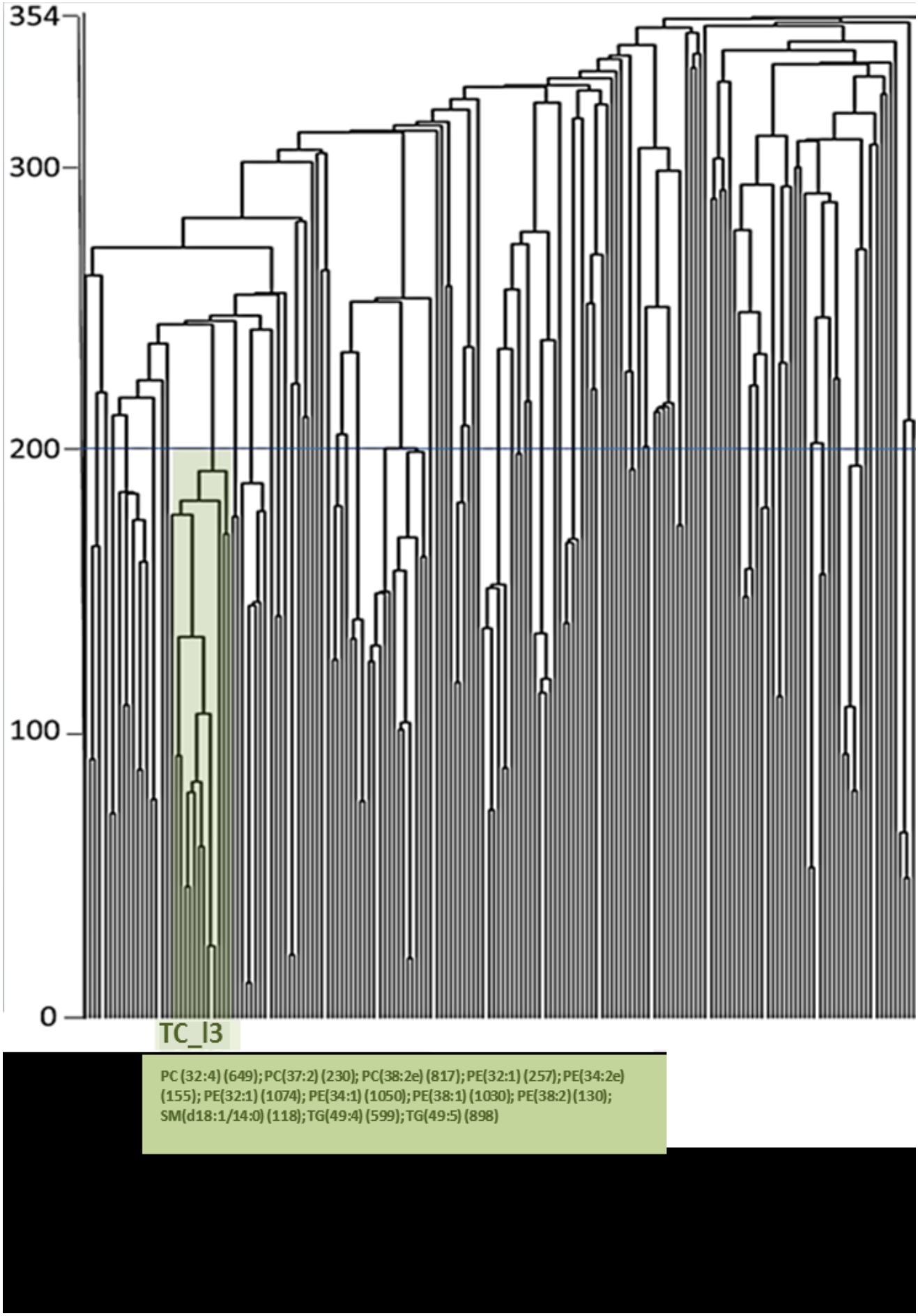


FIGURE 18: Scree plot of explained variances by the TC_ls generated by a TT on serum lipid variables

Consideration of the scree plot (FIGURE 18) indicated that the optimal number of extracted TC_ls is three, because the first three TC_ls explain large parts of the variance whereas a fourth TC_l would not substantially increase the proportion of explained variance.


354
300
200
100
0
TC_I3
PC (32:4) (649); PC(37:2) (230); PC(38:2e) (817); PE(32:1) (257); PE(34:2e) (155); PE(32:1) (1074); PE(34:1) (1050); PE(38:1) (1030); PE(38:2) (130); SM(d18:1/14:0) (118); TG(49:4) (599); TG(49:5) (898)

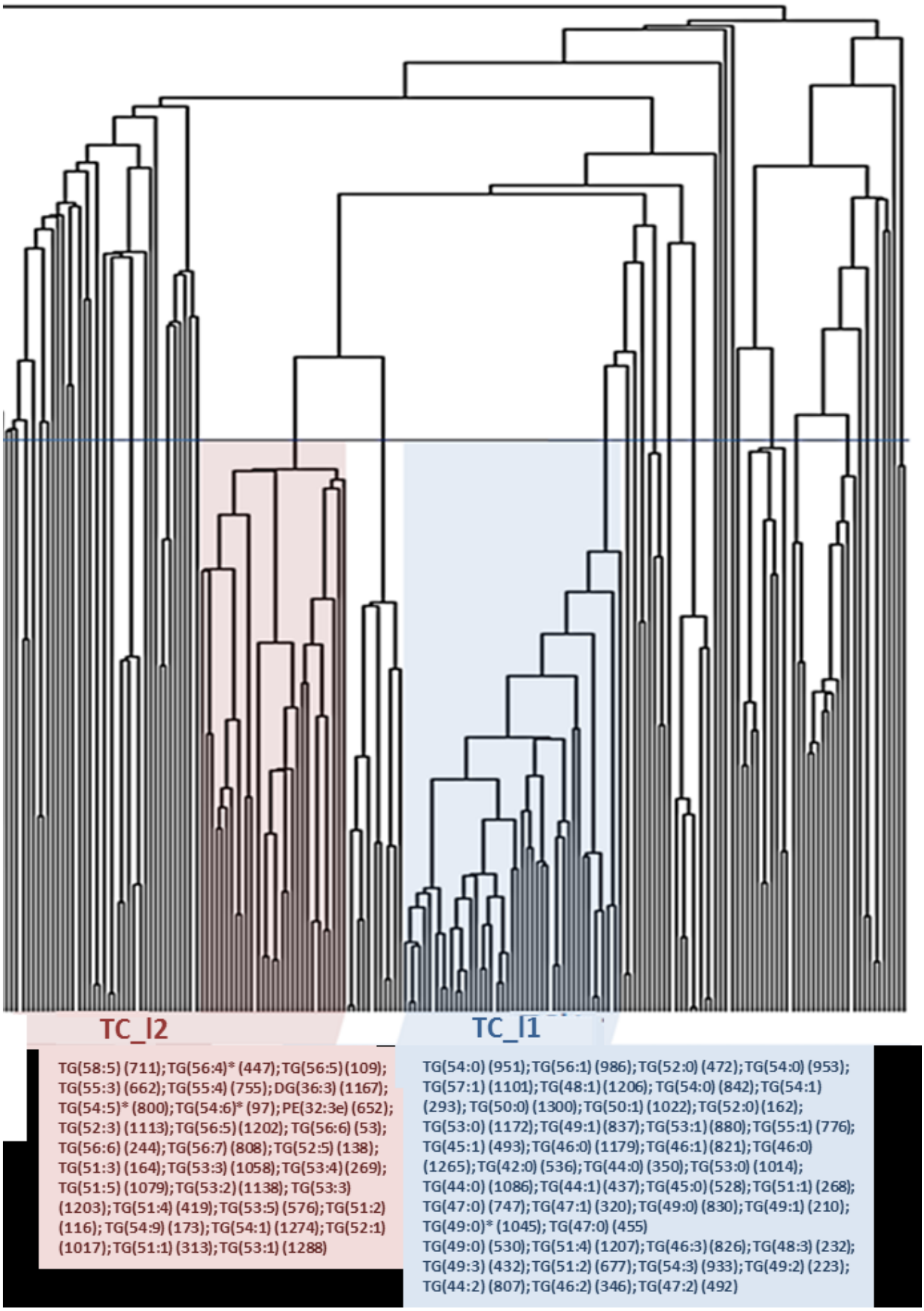


FIGURE 19: Dendrogram resulting from a TT analysis on 353 serum lipid variables; emphasized with colours are the three retained components TC_l1 (blue), TC_l2 (red) and TC_l3 (green)

TABLE 13: Characteristics and loading patterns of three TC_ls generated by TT on 353 serum lipid variables

	TC_l1	TC_l2	TC_l3
Explained variance [%]	9.6	5.4	2.8
Nb of loading variables	43	29	13
Name	Saturated and mono-unsaturated TGs	Polyun-saturated TGs	PCs and PEs
	loading patterns		
Lipid Name (variable nb)	**TC_l1**	**TC_l2**	**TC_l3**
TG(54:0) (951)	0.141		
TG(56:1) (986)	0.141		
TG(52:0) (472)	0.143		
TG(54:0) (953)	0.143		
TG(57:1) (1101)	0.144		
TG(48:1) (1206)	0.146		
TG(54:0) (842)	0.146		
TG(54:1) (293)	0.146		
TG(50:0) (1300)	0.148		
TG(50:1) (1022)	0.148		
TG(52:0) (162)	0.149		
TG(53:0) (1172)	0.149		
TG(49:1) (837)	0.149		
TG(53:1) (880)	0.154		
TG(55:1) (776)	0.154		
TG(45:1) (493)	0.156		
TG(46:0) (1179)	0.156		
TG(46:1) (821)	0.157		
TG(46:0) (1265)	0.157		
TG(42:0) (536)	0.158		
TG(44:0) (350)	0.158		
TG(53:0) (1014)	0.159		
TG(44:0) (1086)	0.159		
TG(44:1) (437)	0.159		
TG(45:0) (528)	0.160		
TG(51:1) (268)	0.160		
TG(47:0) (747)	0.160		
TG(47:1) (320)	0.160		
TG(49:0) (830)	0.161		
TG(49:1) (210)	0.161		
TG(49:0)* (1045)	0.161		
TG(47:0) (455)	0.162		
TG(49:0) (530)	0.162		
TG(51:4) (1207)	0.139		
TG(46:3) (826)	0.142		
TG(48:3) (232)	0.142		
TG(49:3) (432)	0.145		
TG(51:2) (677)	0.146		
TG(54:3) (933)	0.149		

TG(49:2) (223)	0.149		
TG(44:2) (807)	0.156		
TG(46:2) (346)	0.157		
TG(47:2) (492)	0.159		
TG(58:5) (711)		0.161	
TG(56:4)* (447)		0.162	
TG(56:5) (109)		0.173	
TG(55:3) (662)		0.174	
TG(55:4) (755)		0.174	
DG(36:3) (1167)		0.175	
TG(54:5)* (800)		0.177	
TG(54:6)* (97)		0.177	
PE(32:3e) (652)		0.179	
TG(52:3) (1113)		0.182	
TG(56:5) (1202)		0.188	
TG(56:6) (53)		0.188	
TG(56:6) (244)		0.188	
TG(56:7) (808)		0.188	
TG(52:5) (138)		0.190	
TG(51:3) (164)		0.191	
TG(53:3) (1058)		0.191	
TG(53:4) (269)		0.191	
TG(51:5) (1079)		0.193	
TG(53:2) (1138)		0.193	
TG(53:3) (1203)		0.193	
TG(51:4) (419)		0.195	
TG(53:5) (576)		0.195	
TG(51:2) (116)		0.198	
TG(54:9) (173)		0.198	
TG(54:1) (1274)		0.182	
TG(52:1) (1017)		0.185	
TG(51:1) (313)		0.198	
TG(53:1) (1288)		0.198	
PC(32:4) (649)			0.258
PC(37:2) (230)			0.273
PC(38:2e) (817)			0.273
PE(38:1) (1030)			0.290
PE(38:2) (130)			0.290
PE(32:1) (1074)			0.290
PE(34:1) (1050)			0.293
PE(34:2) (1335)			0.293
PE(32:1) (257)			0.298
PE(34:2e) (155)			0.298
TG(49:4) (599)			0.244
TG(49:5) (898)			0.244
SM(d18:1/14:0) (118)			0.254

The stability and consistency of lipid TCs was also assessed by bootstrap methods and by conducting the analogue TT on ±1 to ±3 cut-levels. There was a higher diversity among the sign patterns reported in the sub samples and resulting in lower relative abundance. Thus, all sign patterns that occur in >5% of all sub samples were taken into account. TABLE 14 shows the results from two stability runs. Nine sign patterns occurred in >5% of all subsamples in the first run and eight in the second run, indicating that TC_l2 is the most unstable component and TC_l3 the most consistent one.

TABLE 14: Results of the stability analysis of the TT on 353 serum lipid variables retaining three components on a cut-level of 200, bold are the sign patterns that match exactly to the TC_ls and their frequencies

TC	Similar sign patterns	Frequency
	first stability analysis	
l1	1(+2), 2 (+6), 3(+9), **4**, 5(+4)	13, 9, 10, **6**, 6 = 44%
l2	6(-1)	8%
l3	8(-1), **9**, 11(-2)	7, **53**, 7 =67%
	second stability analysis	
l1	1(+9), 2(+11), 3(+4), 4(+2)	14, 9, 14, 17 = 54%
l2	5 (-1)	6%
l3	7(-1), **9**, 10(-2)	7, **51**, 7 = 65%

In both stability analyses there were three sign patterns namely, sign pattern 7, 10 and 12 in the first analysis and sign pattern 6, 8 and 11 in the second analysis, respectively, that were not similar to any of the three retained TCs from the original analysis.

Further, consistency assessment was performed by a repeated TT on cut-level 197, 198, 199, 201, 202 and 203 to investigate, whether the choice of cut-level has an influence on the structure of the TCs. At all cut-levels the results were exactly the same as for the initial one. The first three components explained 17.8% of the variance and original variables loading on the TCs also did not differ.

3.9 Principal component analysis on identified serum lipids

Analogous to the analyses on spmets, a PCA on the data set of identified serum lipids was performed in addition to TT to reduce dimension in data.

A scree plot (FIGURE 20) indicated three factors to be the optimal number of extracted factors, since a third factor added a quite high proportion of explained variance, while a fourth factor would not substantially increase the proportion of explained variance.

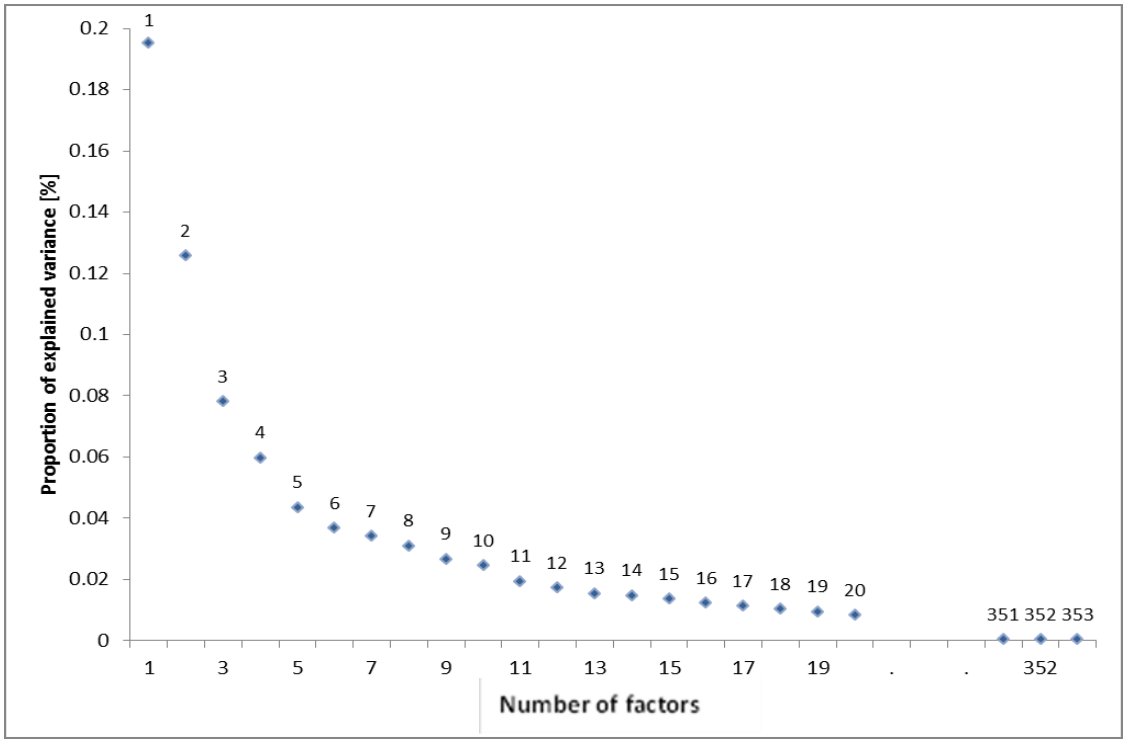


FIGURE 20: Scree plot of the proportion of explained variance for all factors resulting from PCA on identified serum lipids (p=353)

The three extracted factors altogether explained 39.9% of the variance within the lipid data. After orthogonal rotation, factor_l1 explained 19.5%, and factor_l2 and l3 explained 12.6% and 7.8% of the variance. Factor_l1 to factor_l3 were loaded by 132, 124 and 78 original variables with values ≥|0.35| (TABLE 15).

TABLE 15: Description of three factors extracted by a PCA on 353 identified serum lipids; TGs- triglycerides, PEs- phosphatidylethanolamines, PCs- phosphatidylcholines.

	Factor_l1	Factor_l2	Factor_l3
Proportion of explained variance [%]	19.5	12.6	7.8
Number of loading original variables	132	124	78
Name	Saturated and monounsaturated TGs	PEs and PCs	Polyun-saturated TGs
Number of loading variables that load in at least one more factor	26	17	11
	loading patterns		
lipid variable (variable number)	**Factor_l1**	**Factor_l2**	**Factor_l3**
TG(47:1) (320)	0.92		
TG(47:2) (492)	0.92		
TG(48:1) (1206)	0.91		
TG(49:1) (210)	0.91		
TG(49:0) (830)	0.91		
TG(46:0) (1179)	0.9		
TG(46:0) (1265)	0.9		
TG(51:1) (268)	0.9		
TG(47:0) (747)	0.9		
TG(49:2) (223)	0.89		
TG(46:2) (346)	0.89		
TG(46:1) (821)	0.88		
TG(49:0)* (1045)	0.87		
TG(45:1) (493)	0.87		
TG(49:0) (530)	0.87		
TG(49:1) (837)	0.87		
TG(53:0) (1172)	0.86		
TG(47:0) (455)	0.86		
TG(55:1) (776)	0.86		
TG(53:1) (880)	0.86		
TG(52:0) (162)	0.85		
TG(44:1) (437)	0.85		
TG(51:2) (677)	0.85		
TG(44:0) (1086)	0.84		
TG(44:0) (350)	0.84		
TG(53:0) (1014)	0.83		
TG(50:0) (1300)	0.83		
TG(48:3) (232)	0.83		
TG(49:3) (432)	0.82		
TG(45:0) (528)	0.82		
TG(51:2) (116)	0.81		0.4
TG(54:3) (933)	0.81		
TG(50:1) (1022)	0.8		
TG(51:1) (313)	0.8		
TG(57:1) (1101)	0.79		
TG(54:1) (293)	0.79		
TG(44:2) (807)	0.79		
TG(52:0) (472)	0.78		
TG(54:0) (953)	0.78		
TG(54:0) (842)	0.77		
TG(54:0) (951)	0.77		
TG(56:1) (986)	0.77		
TG(51:4) (1207)	0.76		
PE(36:1) (751)	0.76		
TG(46:3) (826)	0.75		
PE(36:1) (871)	0.75		
TG(42:0) (536)	0.74		
TG(50:5)* (633)	0.72		
TG(53:1) (725)	0.7		
TG(53:1) (1288)	0.69		
TG(51:3) (164)	0.68		
TG(54:9) (173)	0.68		
PE(36:1) (823)	0.66		
TG(56:1) (915)	0.66		
TG(42:1) (803)	0.65		
PE(32:3e) (652)	0.63		
PE(34:1) (841)	0.63		
TG(52:3) (666)	0.61		
TG(52:1) (1017)	0.6		0.54
PE(36:1) (145)	0.59	0.36	
LysoPC(14:0) (580)	0.59		
TG(42:2) (827)	0.59		
PE(34:2) (1334)	0.58		

TG(53:3) (1203)	0.56		0.63
TG(57:2) (885)	0.56		
TG(56:2) (474)	0.55		
TG(53:2) (1138)	0.53		0.61
PE(40:8e) (123)	0.53	0.52	
TG(56:2) (546)	0.53		
TG(52:7) (785)	0.52		
TG(56:1) (1192)	0.51		0.4
PC(36:8e) (149)	0.51		
PE(40:8e) (151)	0.51	0.54	
TG(56:2) (1052)	0.5		0.53
TG(50:0) (1184)	0.5		
TG(56:3) (258)	0.5		
PC(37:2) (308)	0.5	0.54	
TG(56:4)* (447)	0.49		
TG(55:1) (870)	0.49		
PE(40:7) (225)	0.48	0.39	
TG(55:3) (662)	0.48		
PC(38:4) (80)	0.48		
TG(58:2) (894)	0.48		
PC(40:3e) (564)	0.47		
TG(54:5)* (800)	0.47		
PE(38:7e) (375)	0.46	0.46	
TG(51:4) (419)	0.46		
PE(36:4) (510)	0.46		
PC(32:2) (148)	0.45	0.35	
PC(38:2) (1123)	0.44		
PE(38:4) (1208)	0.44		0.39
PE(38:4) (243)	0.43		
TG(51:5) (287)	0.43		
TG(52:6) (423)	0.43		
LysoPE(18:0) (545)	0.43		
PC(40:4) (167)	0.42		
PC(38:4) (1025)	0.41		
DG(36:3) (1167)	0.4		0.67
PE(39:9e) (127)	0.4		
PC(33:2)+PE(36:2) (110)	0.39	0.58	
LysoPC(16:1) (355)	0.38		
PE(36:4) (630)	0.37		
PC(36:4e) (1065)	0.36	0.5	
SM(d18:1/22:0) (186)	0.36	0.4	
PE(40:3) (299)	0.36	0.38	
PC(34:3) (89)	0.36		
TG(56:4) (261)	0.35		
TG(56:5) (828)	0.35		
PE(40:7e) (265)	-0.36		
ChoE(18:3) (177)	-0.37	0.4	
PC(38:5e) (190)	-0.37	0.44	
TG(52:2) (542)	-0.37		
SM(d18:1/26:2) (642)	-0.38		
PE(36:5) (464)	-0.4	-0.38	
TG(59:7) (1190)	-0.42		0.68
PE(36:4) (1272)	-0.42		
SM(d18:1/18:2) (583)	-0.42		
PE(40:8e) (609)	-0.42		
PC(40:6) (1033)	-0.43	0.46	
PE(40:6) (386)	-0.47		
TG(57:4) (964)	-0.47		
TG(59:5) (548)	-0.48		
PE(40:6) (217)	-0.49	0.41	
TG(58:6) (310)	-0.54		
TG(59:6) (1067)	-0.56		0.57
TG(55:4) (406)	-0.6		
TG(55:3) (384)	-0.62		
TG(55:2) (359)	-0.7		
TG(56:4) (1227)	-0.73		
TG(58:10)* (1250)	-0.73		
TG(52:4) (482)	-0.79		
TG(50:2) (335)	-0.84		
PC(36:3) (44)		0.82	
PC(30:1) (1019)		0.79	
PE(34:1) (1050)		0.79	
PC(31:1)+PE(34:1) (174)		0.79	
PC(38:2e) (817)		0.79	
PC(37:2) (230)		0.78	
TG(46:4) (57)		0.77	
TG(47:4) (100)		0.76	
PE(34:1e) (152)		0.75	
PC(30:0) (1042)		0.74	
PC(32:4) (649)		0.74	
PE(34:2) (1335)		0.72	
PC(38:7e) (1135)		0.71	
PE(34:2e) (133)		0.71	
PE(36:0) (513)		0.71	

PC(33:0) (912)		0.71	
PE(36:1) (92)		0.71	
PC(40:3e) (415)		0.7	
PC(36:5) (1127)		0.69	
PE(34:2e) (155)		0.69	
PE(34:0) (185)		0.69	
PE(38:1) (1030)		0.68	
PC(38:5) (254)		0.68	
TG(48:5) (1173)		0.67	
PE(38:2) (130)		0.67	
PE(34:0) (1246)		0.66	
PE(32:1) (257)		0.66	
PE(32:0) (646)		0.65	-0.36
TG(48:4) (84)		0.65	
PC(34:3) (156)		0.64	
PE(37:8e) (357)		0.64	
SM(d18:1/16:1) (68)		0.64	
PE(32:1) (1074)		0.63	
PE(38:2) (222)		0.63	
PC(34:3e)+PE(37:3e) (55)		0.62	
TG(46:3) (875)		0.62	
TG(49:5) (898)		0.62	
SM(d18:1/14:0) (118)		0.61	
SM(d18:1/20:0) (52)		0.61	
TG(49:4) (599)		0.61	
TG(48:4) (872)		0.61	
PE(38:3) (189)		0.6	
PC(40:4e) (297)		0.6	
PE(32:1) (713)		0.6	
PC(40:4e) (400)		0.59	
PC(35:2) (78)		0.58	
PC(32:0) (69)		0.57	
PC(37:4)/PE(40:4) (73)		0.57	
TG(44:1) (732)		0.56	
PC(38:3e) (1044)		0.54	
PC(38:2) (111)		0.54	
PC(40:10) (1193)		0.54	
PC(40:5e) (231)		0.54	
TG(52:2) (542)		0.54	
SM(d18:1/25:2) (377)		0.53	
PC(36:5e)+PE(38:5e) (45)		0.53	
PC(36:4) (1036)		0.52	
PC(38:3e) (1128)		0.52	
PE(38:2) (1286)		0.52	
ChoE(16:0) (176)		0.52	
PC(36:4e) (262)		0.52	
TG(53:7) (754)		0.52	
PC(32:3) (175)		0.51	
PC(34:1) (431)		0.5	
PE(40:7e) (518)		0.5	
PC(32:2) (281)		0.49	
TG(51:7) (389)		0.49	
PE(36:2e) (408)		0.49	
Cer(d18:1/16:0) (707)		0.49	
PE(40:4) (819)		0.48	
PE(36:3e) (279)		0.47	
PC(38:5e) (41)		0.47	
PE(34:2) (723)		0.47	
TG(46:2) (1095)		0.46	
LysoPC(18:0) (58)		0.45	
TG(51:6) (640)		0.45	
PC(40:5e) (742)		0.45	
TG(49:2) (537)		0.43	
PE(37:7e) (126)		0.42	
PE(30:0e) (574)		0.42	
SM(d18:1/18:0) (61)		0.42	
SM(d18:1/25:3) (670)		0.42	
PC(38:8e) (114)		0.41	
PC(38:5) (18)		0.41	
PC(40:7) (199)		0.41	
PE(34:2e) (541)		0.41	
TG(46:3) (695)		0.41	
PE(40:5) (812)		0.41	
PE(40:4) (120)		0.4	
PC(34:4) (1156)		0.39	
TG(53:8) (205)		0.39	
PC(40:4e) (1185)		0.38	
PE(40:8e) (560)		0.38	
SM(d18:1/27:3) (890)		0.38	
PC(36:8e) (107)		0.37	
SM(d18:1/18:1) (1092)		0.37	
PE(38:4) (1223)		0.37	
PE(36:4e) (405)		0.37	
TG(46:1) (717)		0.37	

PC(38:5) (802)		0.37	
TG(53:7) (1186)		0.36	
Cer(d18:1/22:1) (965)		0.36	
PC(30:3) (342)		0.35	
LysoPE(18:0) (545)		0.35	
PC(36:5) (79)		-0.36	
Cer(d18:1/22:6) (106)		-0.37	
PE(34:2e) (600)		-0.47	
SM(d18:1/16:1) (434)		-0.54	
TG(56:6) (53)			0.86
TG(56:7) (808)			0.86
TG(56:5) (1202)			0.85
TG(56:6) (244)			0.83
TG(54:6)* (97)			0.81
TG(58:7) (1031)			0.80
TG(56:8)* (1220)			0.79
TG(58:5) (711)			0.78
TG(58:6) (1069)			0.76
TG(56:5) (109)			0.76
TG(58:10)* (838)			0.76
TG(54:4) (1269)			0.70
TG(56:9) (861)			0.70
TG(52:3) (1113)			0.69
TG(56:5) (828)			0.69
TG(51:4) (1247)			0.68
TG(54:7) (216)			0.67
TG(54:5)* (800)			0.65
TG(56:4)* (447)			0.64
TG(53:5) (576)			0.62
TG(58:6) (1057)			0.61
TG(52:5) (138)			0.61
TG(53:4) (269)			0.61
TG(52:6) (423)			0.61
TG(55:7) (956)			0.61
TG(59:5) (548)			0.57
TG(53:3) (1058)			0.56
TG(59:5) (1063)			0.56
TG(56:4) (261)			0.55
TG(55:6) (635)			0.55
TG(55:4) (755)			0.55
TG(51:4) (419)			0.54
TG(51:5) (1079)			0.53
TG(53:1) (1288)			0.53
TG(54:9) (173)			0.53
TG(56:3) (258)			0.53
TG(57:4) (1353)			0.52
TG(55:3) (662)			0.52
TG(58:5) (958)			0.51
TG(55:5) (1008)			0.50
TG(58:6) (310)			0.50
TG(52:7) (785)			0.48
TG(57:2) (885)			0.48
TG(51:3) (164)			0.47
PC(38:5) (64)			0.46
TG(53:3) (1075)			0.45
PC(36:3) (93)			0.45
PC(38:6) (414)			0.44
PE(32:3e) (652)			0.44
TG(57:4) (964)			0.44
TG(58:4) (969)			0.43
PC(40:10e) (1028)			0.42
SM(d18:1/24:2) (251)			0.42
TG(58:3) (918)			0.42
PC(36:5) (1018)			0.41
TG(54:1) (1274)			0.41
PE(38:4) (243)			0.41
TG(56:2) (474)			0.41
TG(58:2) (894)			0.41
TG(48:4) (1070)			0.39
TG(51:1) (313)			0.39
PC(40:7) (115)			0.38
TG(52:0) (162)			0.37
PC(40:9) (124)			0.36
TG(55:1) (870)			0.36
TG(50:0) (1300)			0.35
TG(50:5)* (633)			0.35

Factor_l1 was loaded ≥|0.35| by 132 original variables with most of the very high loading variables being saturated and monounsaturated TGs (TABLE 15) describing the factor best. 91 out of the 124 high loading variables for factor_l2 were (Lyso-) PEs and (Lyso-) PCs, therefore, this factor was designated PEs and PCs. Almost 75% (57 out of 78) of the original variables loading higher than |0.35| on factor_l3 were polyunsaturated TGs.

26 of the lipid variables loading on factor_l1 also occurred in factor_l2 or factor_l3 with a loading of ≥|0.35| (TABLE 15). Further for factor_l2 and l3 about 14% of the loading variables occurred also in at least one of the other factors. This constellation led to a complex structure of the retained factors.

TABLE 16: Results of stability analysis of the PCA regarding the identified serum lipids

Factor	Loaded by original variables [number]	Original variables in bootstrap PCAs [number]	Frequency of original variables in bootstrap PCAs with loading in the same range (75 - 125%) [mean (min – max)]
l1	132	214	82.3 % (20 - 100%)
l2	124	175	83.6% (20 - 100%)
l3	78	150	74.1% (30 - 100%)

To elucidate the stability of factors, again bootstrap methods were used, whereby PCAs with the same characteristics as the initial one were conducted in bootstrap-samples with a size of 80% of the original data set (n=181). Much more original variables loaded on the bootstrap PCA factors than on the initial factors (TABLE 16). Especially in factor_l3 of the bootstrap PCA nearly twice as many loading variables occurred compared to the initial factor. Overall, the factors on lipid variables seemed to be slightly more stable than the metabolite factors (TABLE 16 and TABLE 8). The initial loading original variables occurred in at least 20% of the bootstrap PCA factors with a similar loading, but some of them even in all bootstrap PCA factors. On average there was a frequency of 80% of the original loading variables in the bootstrap PCA factors.

3.10 Comparison of *treelet transform* and principal component analysis on identified serum lipids

Both, TT and PCA, on serum lipid variables generated three components and factors, respectively, summarising the data according to inter-correlations and explaining substantial proportions of the variance within the data (TABLE 17).

TABLE 17: Comparison of TT and PCA on identified serum lipids;

	TT	PCA
Explained variance [%]	17.8	39.8
Components/factors extracted [number]	3	3
Component/factor (number of (high) loading variables)		
Saturated and monounsaturated TGs	TC_l1 (43)	Factor_l1 (132)
Polyunsaturated TGs	TC_l2 (29)	Factor_l3 (78)
(Lyso-)PEs and (Lyso-)PCs	TC_l3 (13)	Factor_l2 (124)
Stability		
Frequency in bootstrap investigation* [mean, min – max]	40.7% (6 – 67%)	80% (74 – 84%)

*** the frequency of similar sing patterns in TC from bootstrap-samples is determined for TT; for PCA it is the frequency of initially loading variables in bootstrap PCA factors**

Again the factors were loaded highly by much more original variables than the TCs. Factor_l1 was loaded by more than three times as many variables as TC_l1. The most extreme ratio could be found for TC_l3 and factor_l2, both characterised by PEs and PCs. The component was loaded by 13 variables whereas the factor was loaded $\geq|0.35|$ by 124 - nearly ten times as many variables (TABLE 17).

With little differences in the order the TCs generated by TT resemble the factors resulting from PCA. Components derived by TT and factors resulting from PCA on a data set of serum lipid variables were quite similar. All saturated and monounsaturated triglycerides loading in TC_l1 could also be found in factor_l1 (FIGURE 21). The same was true for the other original variables loading in TC_l1.

All lipids loading on TC_l2, mostly being polyunsaturated TGs, also loaded variables on factor_l3 (FIGURE 22)

Further, FIGURE 23 shows that the structure of TC_l3 and factor_l2 was comparable since they were both loaded mostly by PEs and PCs.

Regarding the stability, similar sign patterns of TCs from TT showed lower frequency values compared to the initially loading variables from the PCA in bootstrap investigations. TCs with similar sign patterns occurred in 41% of the bootstrap TT whereas variables that loaded on

original PCA factors on average load from 74% to 84% on the bootstrap PCA factors with a similar loading (±25%).

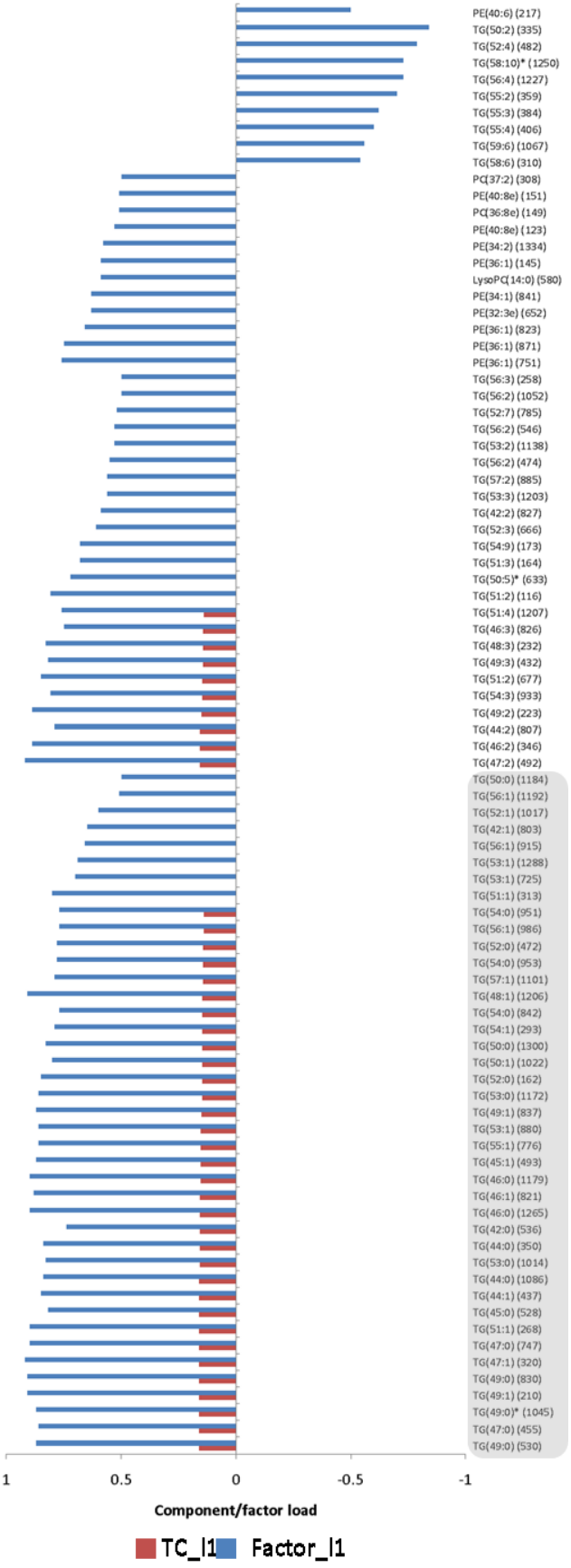


FIGURE 21: Comparison of component/factor loads of TC_l1 and factor_l1, the saturated and monounsaturated TGs component/factor, shaded in grey are saturated and monounsaturated TGs, for reasons of readability only variables loading higher than|0.5| on factor_l1 are displayed

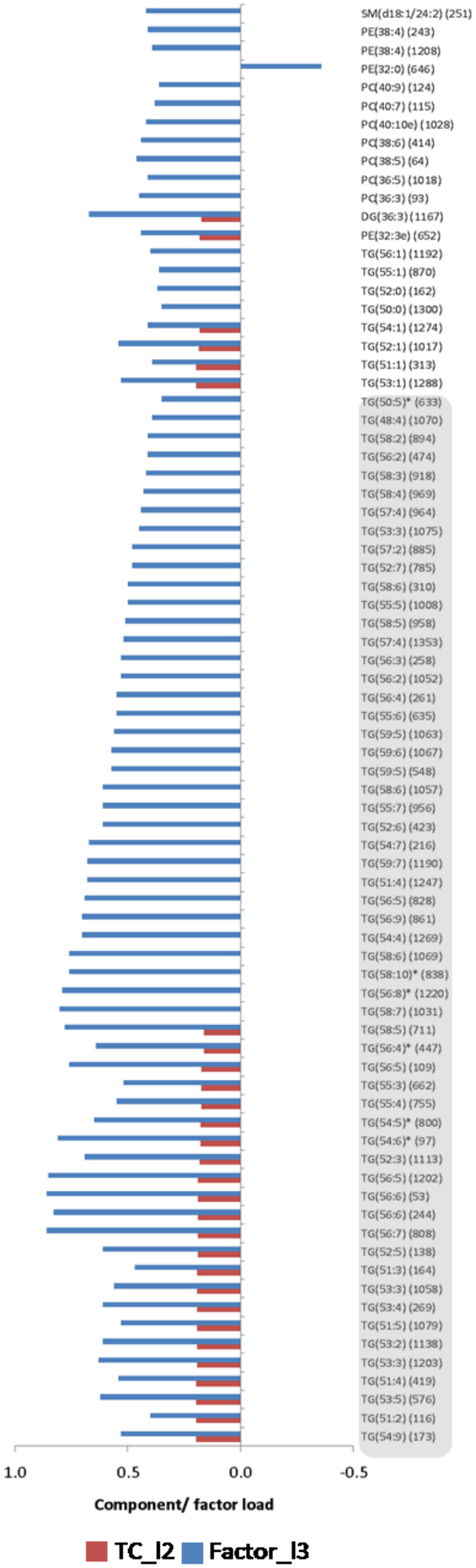


FIGURE 22: Comparison of component/ factor loads of TC_l2 and factor_l3, the polyunsaturated TGs component/factor, shaded in grey are polyunsaturated TGs

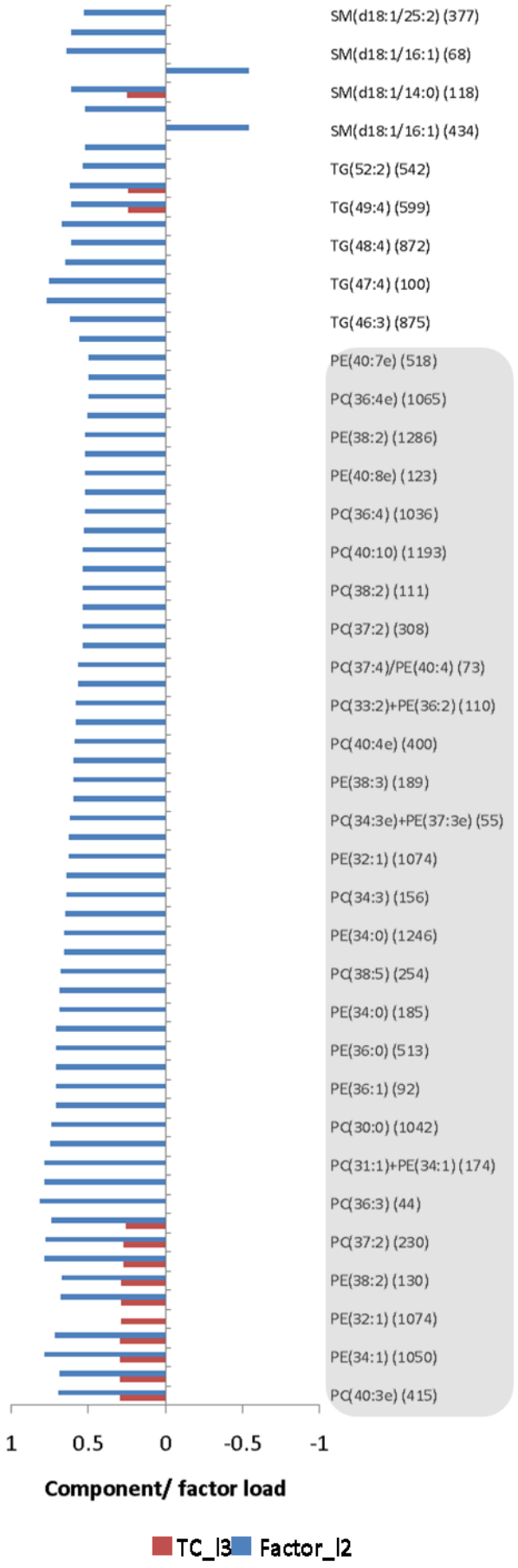


FIGURE 23: Comparison of component/ factor loads of TC_I3 and factor_I2, PEs and PCs component/factor, shaded in grey are (Lyso-)PEs and (Lyso-)PCs, for reasons of readability only variables loading higher than |0.5| on factor_I2 are displayed

3.11 Association of lipid treelet components and factors with anthropometry

All lipid components and factors were correlated to measures of obesity. Since the level of serum lipids depends on anthropometry with heavier individuals having higher levels of serum lipids, the correlations were adjusted for age and sex, and additionally for total serum cholesterol (TABLE 18).

TABLE 18: Partial correlations (corrected for age, sex and total serum cholesterol) between lipid components and factors generated by TT and PCA, respectively, and anthropometry

	Treelet component/ PCA factor	Weight [kg]	Waist circumference [cm]	Hip circumference [cm]	BMI [kg/m²]	WHR	Body fat mass [%]	WHtR	BAI	visceral fat [cm³]	subcutaneous fat [cm³]	abdominal total fat [cm³]	visceral/abdominal subcutaneous fat ratio
Saturated and mono-unsaturated TGs	TC_l1	0.136	0.172		0.170	0.197	0.159	0.185	0.109	0.147	0.161	0.146	0.103
	Factor_l1	0.103	0.139		0.134	0.174	0.135	0.152		0.139	0.143	0.125	0.103
Polyun-saturated TGs	TC_l2	0.232	0.264	0.211	0.272	0.212	0.239	0.273	0.215	0.262	0.273	0.276	
	Factor_l3	0.292	0.328	0.283	0.326	0.235	0.232	0.331	0.274	0.306	0.299	0.328	
PCs and PEs	TC_l3	-0.116	-0.154	-0.151				-0.137	-0.113	-0.239	-0.104	-0.159	-0.184
	Factor_l2	-0.175	-0.219	-0.161	-0.172	-0.186	-0.114	-0.213	-0.140	-0.297	-0.137	-0.198	-0.226
Degree of correlation		<= -0.2	-0.2 to - 0.13		-0.129 to - 0.1		0.1 to 0.129	0.13 to 0.2		>= 0.2			

The TC_l and the factor_l consisting mostly of saturated and monounsaturated TGs, TC_l2 and factor_l1, were positively correlated to all measures of anthropometry, except hip circumference. The correlations were all in a moderate ranging from 0.103 to 0.197. Also, all MRT body fat data were associated directly to the TC_l1 and factor_l1. Again the correlations were in a moderate range between 0.10 and 0.15 (TABLE 18). In all cases of common associations of the TC_l and the factor_l the correlation coefficients between the TC_l1 and anthropometry were stronger than that between factor_l1 and anthropometry. The mean and the range of the correlation coefficients for TC_l1 were 0.153 and 0.103 to 0.197 and for factor_l1 they were 0.135 and 0.103 to 0.174, respectively.

Positive associations of the anthropometric parameters and the polyunsaturated TGs component and factor were stronger. For all anthropometric parameters were correlated higher than 0.21, the strongest relation could be found between the polyunsaturated TGs and the WHtR with 0.273 and 0.331, respectively. Also, most of the MRT data were associated positively with TC_l2 and factor_l3; again the correlations were in a fairly strong

positive range between 0.26 and 0.33. Only the ratio of visceral and abdominal subcutaneous fat was not correlated to the TC_l2 and factor_l3.

TC_l3 and factor_l2, the PEs and PCs component and factor, were negatively correlated to weight, waist and hip circumference, WHtR and BAI. Factor_l2 additionally was inversely correlated to BMI, WHR and body fat mass. Even stronger negative correlations could be observed between both, the TC_l3 and the factor_l2, and the MRT body fat data. The correlations were between -0.10 and 0.30, whereby the strongest correlations were observed between the component and the factor and visceral fat with correlation coefficients of -0.239 and -0.297, respectively (TABLE 18).

3.12 Association of lipid treelet components and factors with intestinal microbiota

The factors and TCs generated from serum lipid variables showed lower correlations with intestinal microbiota compared to the factors and TCs from serum metabolites (TABLE 11 and TABLE 19). The factor that was described mainly by saturated and monounsaturated TGs (factor_l1) was correlated to three microbiota bands. For only one band, b300, an intersection with the correlation of the appropriate TC (TC_l1) and microbiota could be found. Factor_l2 and TC_l3 (PEs and PCs) showed correlations for one (b636) and three (b198, b337 and b344) bands, respectively, but none of the associations was with the same DGGE band. This was also true for factor_l3 and TC_l2 (polyunsaturated TGs), where only TC_l2m showed two correlations with a microbiota band (b893 and b921), but factor_l3 was not linked to any of the DGGE bands.

TABLE 19: Correlations of lipid factors and TCs with intestinal microbiota, stated are φ-coefficients for pairs with significant correlation ($p < 0.05$); **bold** are the DGGE bands which are associated with the factors and the TCs; 1*- diversity as total number of DGGE bands per sample and 2**- Shannon-Wiener diversity index

TC or factor \DGGE band	Saturated and monounsaturated TGs		Polyunsaturated TGs			PEs and PCs		
	TC_l1	factor_l1		TC_l2	factor_l3		TC_l3	factor_l2
b300	-0.164	-0.166	b893	0.154		b198	0.163	
b673		0.191	b921	-0.136		b337	-0.167	
b686		-0.150	diversity			b344	0.161	
diversity			1*			b636		0.150
1*			2**			diversity		
2**						1*		
						2**		
Degree of correlation	<-0.2	-0.2 to -0.13	0.13 to 0.2	>0.2				

Overall the φ-coefficients between the lipid factors and TCs and intestinal microbiota were less strong compared to the associations of the metabolite factors and TCs. The absolute values ranged from 0.136 to 0.191 and had a mean value of 0.160.

Regarding the diversity, there was no correlation recognisable with the lipid TCs and factors. No correlation coefficients higher than |0.13| could be observed.

3.13 Sensitivity analysis

Comparison of the participants groups with and without MRT data and information on intestinal microbiota, respectively, showed only minor disparities between the corresponding groups (TABLE 20).

TABLE 20: Sensitivity analysis regarding the participants with missing values on microbiota data and MRT data, respectively

	with data (n= 216)	without data (n=10)	*p**	variable
microbiota	63.29 ± 8.84	69.27 ± 10.59	0.11	age [years]
	45.8% / 54.2%	70% / 30%	0.2	sex (male/female)
	27.30 ± 4.22	27.87 ± 4.50	0.71	BMI [kg/m²]
	0.004 ± 2.58	-0.08 ± 2.23	0.91	score TC_m1
	0.05 ± 5.91	-1.03 ± 4.06	0.44	score TC_l1
	with data (n= 177)	without data (n=49)	*p**	variable
MRT data	62.83 ± 8.68	66.20 ± 9.62	0.03	age [years]
	48.6% / 51.4%	40.8% / 59.2%	0.42	sex (male/female)
	27.16 ± 4.20	27.92 ± 4.31	0.28	BMI [kg/m²]
	0.02 ± 2.50	-0.07 ± 2.78	0.85	score TC_m1
	-0.13 ± 5.84	0.46 ± 5.82	0.53	score TC_l1

**p*-Value from t-test for metric variables (age, BMI, score TC_m1, score TC_l1) and from Fisher's exact test for nominal variable (sex)

Participants with data on microbiota and MRT data, respectively, are slightly younger than those without the data (microbiota: 63.3 vs. 69. 3 years, MRT: 62.8 vs 66.2 years), but only for the participants with and without MRT data this difference was significant (TABLE 20). No significant differences were observed in sex, BMI and the scores of TC_m1 and TC_l1.

4. Discussion

4.1 Discussion of methods

This thesis investigates the usability of a new, unestablished dimension reduction method, the *treelet transform*, compared to an established method, the principal component analysis, within the scope of a complex serum metabolite and serum lipid data set. This approach aimed to investigate a statistical method that generates stable and less complex results in data sets with more variables than observations in which PCA seems to be limited. Both methods were used to generate summarising, latent variables from a large number of serum metabolite and lipid data, respectively, with acceptable loss of information and to associate these summarising variables with anthropometric parameters and variables on intestinal microbiota. Both approaches resulted in comparable comprising variables, but those variables generated by TT explained less variance both for the metabolite and the lipid data, but were more unambiguous regarding the interpretation. This aspect of intuitive interpretation and sparse components is seen as the most important advantage of TT in comparison to established methods considering dimension reduction, such as PCA [78]. Regarding the lower proportion of explained variance, other studies even found that TCs from TT explained almost the same proportion of variance as PCA factors [26] and furthermore the lower proportion of explained variance can be made reasonable by an aspect which resulted in better interpretability of the latent variables, TCs are loaded by considerably less original variables than correspondent PCA factors. Ascertained stability for TCs was marginally lower than that observed in other studies [26] but similar in range and it was also comparable or even stronger than the stability of PCA factors. As a result of the investigated characteristics TT seems to represent an attractive and auspicious alternative to PCA for the analysis of high-dimensional data [78].

4.1.1 Discussion of data assessment methods

The blood utilized for metabolite and lipid analysis was taken from the participants only at one time point, but since the design of the study was cross-sectional and did not aim to investigate aspects of temporality this was not seen as an impairment. In general the blood draw followed a standardised procedure. What has to be considered is that not all participants provided a fasting blood sample and some of them were non-fasted.

Also the anthropometric parameters were assessed according to a standardised protocol. For all participants the same devices were used and only a limited number of study personnel performed the measurements. This facilitates a consistent assessment of anthropometry and reduces the risk of measurement errors. The utilization of different parameters for similar characteristics, such as body fat mass by skinfold thickness and MRT

data, also minimises the risk of the exclusive application of incorrectly measured parameters. The additional application of MRT data is also beneficial since a validation study showed partly only moderate correlations of measured anthropometric parameters and body fat variables assessed by MRT analysis [79].

A favourable method to determine small polar metabolites and lipids in the blood samples is the untargeted metabolomics approach used in the present study. With this method it is possible to determine not only a pre-defined set of compounds but all existing ones; thereby it enables to mirror the complex system of human metabolism in a better way [80]. Another beneficial issue considering the blood sampling was the circumstance that the serum fraction of the blood samples was used for metabolomic analysis, since in serum samples metabolites can be assessed with higher reliability than in plasma samples [81].

4.1.2 Strength and limitations of study design

The present investigation has its main strength in the study design. It was conducted in the framework of a large prospective cohort study and the metabolite data were assessed with an untargeted approach. An advantageous aspect of prospective cohort studies, which was basis for the study population used in this study, is the circumstance that several characteristics which are thought of as exposures and outcomes can be investigated at the same time [82].

The aim of this investigation was to examine serum metabolite and lipid data regarding their correlation to anthropometric parameters either measured manually, such as body weight or hip circumference, or determined by MRT analysis and appropriate algorithms, such as visceral or subcutaneous body fat mass, to obtain an extensive idea on how serum metabolite and serum lipid data are associated to such parameters.

An advantage of the study is that it was conducted in the general population, so it was possible to get data from normal weight participants as well as from overweight or obese participants that are meant to be different according to their metabolite and lipid profiles [83-85]. Also the untargeted metabolomics approach to detect all existing metabolites and lipids in the serum samples is profitable, since it enhances the depiction of the complex human metabolism compared to approaches detecting only a pre-defined set of metabolites.

4.1.3 Strength and limitations of methods

A particular strength of this study is the harmonised way of collecting anthropometric and metabolite information. All participants were examined with the same methods and also the blood sampling followed a standardised procedure. Also the investigation of a lot of compounds using an untargeted metabolomics approach is an advantage since metabolomics become more and more important in different scientific disciplines [86, 87]. The use of an untargeted metabolomics approach is advantageous as it enhances the illustration of the complex human metabolism [80, 88] since it investigates more than already known compounds and it thereby might improve the understanding of pathogenesis.

The amount of metabolite data was summarised by means of two statistical techniques; both considering an inter-correlated structure, which has to be assumed for serum metabolite and serum lipid data. The PCA is an established method, which is used for many investigations, especially within the scope of "–omics" sciences for a long time [89-92]. The second method, TT, is a quite new, rarely applied dimension reduction method. Both techniques reduced the dimension among serum metabolite and serum lipid data to quite adequate summarising variables, which explained a large proportion of the variance within the data.

Since both methods aim to explain large parts of variance it was favourable to exclude variables with very high variances in the present study because it can be assumed that the extremely high variance can be the result of measurement errors [21]. Inclusion of these variables could give rise to misleading results since these variables load highly on components or factors explaining large parts of the variance.

A further strength was the quite new method TT, which was used in this study and compared to a very established one regarding crucial characteristics, namely proportion of explained variance, interpretability of the generated, latent variables, stability and use for further analyses. Thereby, the usability of the new method could be assessed in a proper way.

Additionally, the associations between serum metabolite and lipid variables and anthropometric parameters were calculated with corrections which reflect the actual situation appropriately. It is known that concentrations of serum metabolites and serum lipids depend on age and sex of a person [71, 72]. Therefore, the correlation properly took into account both relevant variables by partial correlation.

The sensitivity analysis regarding the participants with and without data on intestinal microbiota and MRT data excluded that observed correlations between metabolite and lipid components and anthropometry and intestinal microbiota result from confounding considering the probability of providing the information. Since for MRT data as well as for microbiota data participants with missing values had almost the same age, BMI and scores of

TC_m1 and TC_l1 compared to the participants with existent data it was shown that the observed correlations depend on the investigated parameters.

4.2 Discussion of results

4.2.1 Discussion of dimension reduction methods

In the present thesis serum metabolite and lipid data were reduced in their dimension with two methods, the PCA as an established one and TT as a new method. Both methods generated almost the same number of retained summarising variables - a result which was also obtained in previous studies using both methods [26]. Regarding the TT on 121 serum metabolite variables five components were extracted, PCA generated four factors. All original variables loading in TCs and contributing to the interpretation of the components could also be found in one of the factors demonstrating the comparability of both approaches. In each case there was a component characterised by loadings of amino acids and a second component characterised by fatty acids; a factor formation that was also noticeable in previous studies [84, 93].

One advantageous aspect of TT is the improved interpretability of the retained components since only a few original variables loaded on them. In both, investigations of serum metabolites and lipids, respectively, more than twice as many original variables loaded $\geq|0.35|$ on the PCA factor compared to the loadings on the appropriate TC. On average nearly 5 times more variables loaded on the factors compared to the components generated by TT, at least 2.4 times more in the case of TC_m1 and factor_m4. The highest value was observed for TC_l3 and factor_l2 with 9.5 times more original variables loading on the factor compared to the TC. Another advantage of TT regarding the interpretation of retained, latent variables is that original variables do either load or do not load on the component. Corresponding to the structure of the cluster tree and the chosen cut-level an original variable is part of a component or not. No decision had to be made which loading indicates that an original variable is relevant for the interpretation of the component; a decision that can influence the structure of factors from a PCA.

With this aspect the issue of lower proportion of explained variance by components retained from TT coincides. In the present investigation, the PCA factors explain about twice as much variance as the components from TT. This fact has to be considered in case of an investigation to generate summarising variables which explain as much variance within the data as possible. But due to the circumstance that less original variables load on TCs it elucidates the reason of the lower proportion of explained variance. However, the overall better interpretability is an advantage of the TT method that cannot be neglected in the investigation of high-dimensional data.

The stability check of TT was more severe due to it takes into account complete components in bootstrap-samples while the PCA stability check only accounts for loading variables of initial factors in bootstrap factors with a similar (±25%) loading. Therefore, it is indeterminable if the PCA factors are more stable than the TCs although the original variables loading on a factor occur more often in bootstrap factors than the TC sign patterns occur in bootstrap components. Stability was moderate in both cases.

Overall the present study proves TT, a method which was recently introduced also in nutritional epidemiology [26, 94-96], to be a complementary dimension reduction method to PCA. By resulting in sparse components, TT offers interpretational advantages with no recognisable disadvantages.

In regard to the progress of biochemical techniques and the metabolomic research overall, in targeted as well as in untargeted approaches, [97] it appears to be useful to further progress the statistical methods to obtain adequate results from advanced biochemical and statistical approaches.

4.2.2 Discussion of observed correlations

The aim of many studies is to identify biomarkers of diseases. However, it is often the case that overweight or obesity increases the disease risk [5, 6, 8, 9]. Therefore, it would be desirable to have metabolomics data for anthropometric parameters that are related to a higher risk of NCD. The present study found some evidence that human anthropometry is associated to serum metabolites and serum lipids.

Other studies also found associations of metabolites and anthropometric parameters, such as BMI [98-100]. Also the negative correlations of PCs and PEs and body fat parameters, such as visceral fat or subcutaneous fat, could be observed in other studies, as well as the positive correlations between TGs and body fat parameters [101]. Amino acids detected in blood samples and body fat parameters showed a positive association in the present study and in previous studies [102-104]. These findings of a link between serum amino acids and obesity measurements suggest that serum amino acids are an indicator of a high energy intake and a positive energy balance, both markers for obesity. Also in another study amino acids were associated with BMI [105]. Another aspect confirming the relationship between amino acids and anthropometry is the observed association of another TC_m and factor_m with anthropometric parameters- the amino acid derivate component and factor. All other metabolite TCs and factors showed no or only scattered associations to anthropometry. This was the case for the TC and the factor highly loaded by fatty acids (TC_m2 and factor_m3) that showed no correlation with any of the anthropometric parameters. This does not conform findings from other studies, where an association of fatty acids with anthropometry was observed [34]. The reason for this contradiction might be the different analytical material from which the fatty acids were determined. Dahm et al. (2011) determined fatty

acids in adipose tissue whereas, in the present study this was done in blood samples. An association of fatty acid concentration to a disease risk influenced by anthropometric parameters, risk for type 2 diabetes, found a study of Kröger et al. [106] in the EPIC-Potsdam cohort. Concentration of some fatty acids in erythrocyte membranes, such as heptadecanoic acid, was negatively associated to the risk of type 2 diabetes whereas some monounsaturated fatty acids increased the disease risk.

In the present study, a range of associations could be detected between the lipid TCs and factors and anthropometric parameters. Almost all parameters, except the ratio of visceral to abdominal subcutaneous fat, were linked to serum polyunsaturated lipids (TC_l2 and factor_l3) with positive correlation coefficients in a range of 0.21 to 0.33. This coincides with correlations observed in other studies [107, 108] and also underlines the potential risk of both, elevated serum triglycerides and obesity markers.

Also the TC and the factor characterised by saturated and monounsaturated TGs (TC_l1 and factor_l1) were positively correlated to almost all anthropometric measurements. The only exception was hip circumference for which no correlation could be observed. The correlation coefficients were compared to those of TC_l2 and factor_l3 and in general quite moderate and were in a range between 0.10 and 0.20. Correlations of triglycerides could be also observed in another study, but here the correlation was sex-specific, only significant in women [109].

Also the latent variables of PCs and PE were quite frequently correlated to anthropometric measurements. While factor_l2 was associated to all parameters, TC_l3 was not correlated to BMI, WHR and body fat mass but to all other parameters. Her, all correlation coefficients were negative and ranged between -0.10 and -0.30. Highest correlations could be observed between component and factor, respectively, and body fat parameters from MRT data. These findings coincidence with observations from another study where a positive correlation between fat free mass of the body and serum PCs could be ascertained [110].

Both, studies in humanised mice [111, 112] and in humans [113, 114], indicate that the microbial community is influenced by the nutrients and metabolites ingested by an individual. But it is also known that the microbiota influences the human metabolite profiles [115-117] and that it is also connected with anthropometry and health [118-122]. The human gut microbiota is known to convert substances obtained by diet, e.g. the bacteria generate short chain fatty acids from ingested carbohydrates and proteins [123, 124]. It also has the reputation to have an influence on the development of metabolic syndrome [125, 126]. Also, the research on microbiota and diseases that are influenced by anthropometric parameters, such as diabetes, showed that especially the intestinal microbiota is linked to the disease [127]. Since short chain fatty acids provide a significant proportion of daily calories [128, 129], there is also a link to overweight. Previous studies showed that the energy absorption was changed in persons among whom also the gut microbiota was modified [130]. Therefore, it is not astonishing that in the present investigation associations between the metabolites and the intestinal microbiota were detectable. Also, the diversity

of gut microbiota is supposed to have an influence on the human metabolism [131], an assumption that the present study could not confirm by a correlation between the number of detected DGGE bands and some metabolite or lipid factors and components. In the present investigation there was no correlation between serum metabolites or lipids and the microbial diversity. It is possible that the method to assess intestinal microbiota does not assess the diversity in an optimal way; the DGGE method does not determine the bacterial species and other studies combine it with broader sequencing approaches [132-134]. Especially the aspect of microbial diversity is interesting for further research since in some studies a decreased diversity is linked to diseases [135, 136]. Particularly in the research area of microbiota and human metabolism the metabolomics approach entrenches more and more [137] an aspect that also becomes important. It reinforces the fact that the present investigation detected much more links between the serum metabolites and the gut microbiota (78 correlations) compared to the serum lipids and gut microbiota (15 correlations).

4.3 Conclusions and implications for public health

Since insights from epidemiological studies covering metabolomics, especially untargeted metabolomics, become more and more important in understanding disease development and prevention strategies, statistical methods to investigate enormous data sets need to be enhanced and improved. With the TT, the present thesis suggested a useful alternative to an established method, the PCA, concerning the dimension reduction procedure that assists to handle some challenging problems of other methods a deduction that was also made in conclusion to other studies [26].

Since obesity depicts a growing problem within the public health system as it is a risk factor for many chronic diseases, it is important to figure out metabolites, which are associated to overweight and obesity. The investigation of the whole metabolome instead of only single metabolites enables a better understanding of disease development and thereby improves prevention strategies. Identification of obesity related metabolites can help to identify high-risk individuals, which enables a longer life with high quality of life and without any chronic disease. An aspect that gains importance due to the increasing life expectancy in the Western society [138].

Another fact is that many metabolites turn out to be risk factors that are independent from established ones for several chronic diseases, such as type 2 diabetes [19]. Metabolites are in some cases common risk factors for several chronic diseases [139], therefore, it is also obvious that some metabolites are common for obesity and chronic diseases. Therefore, metabolites can be used in addition to classic, traditional risk factors for obesity and disease risk prediction and in some cases assessment of metabolite profiles may be more adequate and feasible. The assessment of risk factors by questionnaires is for example less objective than the measurement of metabolites in blood samples. These assumptions are important for the development of prevention strategies.

In the event of a confirmation of the observed correlations in prospective cohort studies individual metabolomic profiles can in future be used for the identification of persons that have an elevated risk for chronic diseases.

Summary

Anthropometric parameters giving hints on overweight or obesity are in most cases high-risk factors for diseases, especially for chronic diseases, such as cardiovascular disorders or type 2 diabetes. To achieve a better understanding of the disease pathogenesis it is necessary to get a complex insight into human metabolism. Additionally, the identification of serum metabolites and lipids being associated with overweight could help to detect high-risk persons.

Numerous biochemical developments, particularly *untargeted metabolomics* due to the high amount of detectable metabolite compounds, provide promising approaches. However, the arising comprehensive data sets require further development in the field of analysis, since established methods, such as principal component analysis, hit the wall dealing with such data sets and generate instable results that are difficult to interpret. In many cases it is advantageous to reduce the original metabolite data in their number of variables without a loss of information. In regard to that point, principal component analysis has established in many epidemiological studies. Nevertheless, if the data set contains more variables than observations principal component analysis generates instable results. Furthermore, the extracted factors are difficult to interpret, because every factor is loaded by all original variables and a decision has to be made which load indicates relevance of an original variable for the factor explanation. On behalf of this scenario that many variables have to be summarised further dimension reduction method are investigated. Regarding this demand, the present thesis compares the use of a principal component analysis with a new less established method, the *treelet transform* (TT), considering several characteristics. In the frame of a sub-study (n=226) in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam cohort, variables which describe serum metabolites and serum lipids were reduced with both methods. In addition to this methodological comparison the generated summarising variables were investigated regarding their association to anthropometric parameters. Anthropometry was consistently assessed in this study. From all participants, body weight, body height and circumferences as well as thickness of skin folds were measured to calculate BMI, WHR BAI and body fat mass. Additionally, MRT data which were assessed in the majority of the participants, were used to determine further body fat parameters.

TT was based on the inter-correlation structure among the original variables. By means of the inter-correlation structure a cluster dendrogram was constructed, that always groups the highly correlated variables. This dendrogram was cut on an optimal level that was assessed by cross-validation methods. By this, components were generated with variables being in the same cluster at this level. Components that explain largest parts of the variance were extracted and used for further analysis.

Both statistical methods, TT and PCA, resulted in a comparable number of summarising variables. TT summarised the metabolite data to five components, whereas the principal component analysis resulted in four factors. Composition of components and factors was comparable; in every case there was one amino acid, one fatty acid, one amino acid derivate and one sugar compound/carboxylic acid component and factor. Solely the sugar alcohol could be found in only one summarising component.

Investigation of serum lipid variables showed similar results. Both, TT and PCA, generated three components and factors, respectively, in each case one was loaded by saturated or monounsaturated triglycerides, one by polyunsaturated triglycerides and one by (lyso)phosphatidylethanolamines and (lyso)phosphatidylcholines. Despite the similar number of components and factors the factors from the PCA explained in the two cases, serum metabolites and serum lipids, twice as much variance as the components from the TT. The high proportion of explained variance in the PCA could be deduced as a consequence of the fact that all original variables load on every extracted factor with different loadings. In contrast, on the extracted components only these original variables load that have been in a correlated group in the dendrogram on the chosen cut-level. That is why the components became less complex and easier to interpret, but also the proportion of explained variance decreased.

The factors as well as the components showed numerous correlations with anthropometry. Regarding the trend the correlation coefficients were always in the same direction; only with respect to the strength the coefficients between components and anthropometry showed different values compared to the coefficients between factors and anthropometry, whereas both the coefficients with the component and the factor occurred to be higher.

Overall, this investigation confirms TT as an alternative to a PCA, in the case that there are more variables than observations in the data set.

The observed correlations have to be considerably proved in further studies. In case of a confirmation of these correlations they can be used for the identification of high-risk persons regarding chronic diseases.

Zusammenfassung

Anthropometrische Parameter, die Hinweise auf Übergewicht oder Fettleibigkeit geben, sind häufig starke Risikofaktoren für Erkrankungen, vor allem für chronische Erkrankungen wie kardiovaskuläre Erkrankungen oder Typ 2 Diabetes. Um die Pathogenese dieser Erkrankungen besser zu verstehen, sind komplexe Einblicke in den menschlichen Metabolismus notwendig. Zudem kann die Identifizierung von Serummetaboliten und -lipiden, welche mit Übergewicht assoziiert sind, helfen, Hochrisiko-Personen zu erkennen.

Hierfür bieten zahlreiche biochemische Entwicklungen, insbesondere die *untargeted Metabolomics* durch die hohe Anzahl an nachweisbaren Stoffwechselverbindungen, aussichtsreiche Ansätze. Die dadurch entstehenden umfangreichen Datensätze erfordern jedoch ebenfalls Fortschritte im Bereich der Auswertung. So stoßen etablierte statistische Methoden, wie z.B. die Hauptkomponentenanalyse, bei eben solchen Datensätzen an ihre Grenzen und generieren zum Teil instabile und schwer zu interpretierende Ergebnisse. In vielen Fällen ist es für die Ergebnisinterpretation sehr vorteilhaft, die zu Grunde liegenden Metabolitdaten hinreichend in der Anzahl der Variablen zu reduzieren, ohne jedoch einen zu starken Verlust an Information in Kauf nehmen zu müssen. Hinsichtlich dieser Aufgabe hat sich die Hauptkomponentenanalyse in vielen epidemiologischen Studien etabliert. Liegt jedoch ein Datensatz vor, in dem es mehr Variablen als Beobachtungen gibt, erzielt eine Hauptkomponentenanalyse eher instabile Ergebnisse. Die extrahierten Faktoren sind zudem schwer zu interpretieren, da jeder generierte Faktor von allen Originalvariablen geladen wird und entschieden werden muss, ab welcher Ladung eine Variable relevant ist. Für dieses Szenario von zahlreichen Variablen, welche zusammengefasst werden sollen, werden weitere Dimension-Reduktions-Methoden untersucht. Hinsichtlich dieser Anforderung wurde in der vorliegenden Arbeit die Anwendung der Hauptkomponentenanalyse mit einer neuen, weniger etablierten Methode, dem *Treelet Transform* (TT), in Bezug auf verschiedene Charakteristika verglichen. Innerhalb einer Studie an einer Subkohorte (n=226) der European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam-Kohorte wurden hierbei mit beiden Methoden Variablen reduziert, welche Serummetabolite und -lipide beschreiben. Zusätzlich zu dem methodischen Vergleich, wurden die generierten zusammenfassenden Variablen auf ihre Assoziation mit anthropometrischen Parametern untersucht. Die Anthropometrie wurde innerhalb dieser Studie sehr einheitlich erfasst. Alle Probanden wurden hinsichtlich Körpergewicht, Körpergröße und Umfängen sowie Dicke von Hautfalten vermessen, um aus diesen Werten BMI, WHR, BAI und Körperfettmasse zu errechnen. Zusätzlich wurden MRT-Daten, die von einer großen Zahl Probanden vorhanden waren, genutzt, um weitere Körperfettparameter zu ermitteln.

TT basierte auf der Korrelation der Originalvariablen untereinander. Mit Hilfe dieser wurde ein Clusterdendrogramm erstellt, welches die hochkorrelierten Variablen untereinander gruppiert. Dieses Dendrogramm wurde an einer, mit Hilfe von Kreuzvalidierungen

bestimmten, optimalen Höhe geschnitten, wodurch einzelne Komponenten entstanden, die auf dieser Höhe noch einen gemeinsamen Stamm hatten. Die Komponenten mit der höchsten Varianzerklärung wurden extrahiert und für weitere Analysen genutzt. Beide statistische Methoden generierten eine vergleichbare Anzahl an zusammenfassenden Variablen. TT fasste die Metabolitdaten zu fünf Komponenten zusammen, während die Hauptkomponentenanalyse in vier Faktoren resultierte. Komponenten und Faktoren waren in ihrer Zusammensetzung sehr vergleichbar, so gab es für Aminosäuren, Fettsäuren, Aminosäurederivate und für Zuckerverbindungen und Carbonsäuren jeweils eine Komponente und einen Faktor. Lediglich Zuckeralkohole konnten nur als hauptsächlich ladende Variablen einer Komponente gefunden werden. Eine Untersuchung der Serumlipide produzierte sehr ähnliche Ergebnisse. Beide Methoden resultierten in drei Komponenten bzw. Faktoren, jeweils geladen von 1.) gesättigten und einfach ungesättigten Triglyceriden, 2.) von mehrfach ungesättigten Triglyceriden und 3.) von (Lyso-) Phosphatidylethanolminen und (Lyso-) Phosphatidylcholinen. Trotz der in etwa gleichen Anzahl an Komponenten und Faktoren übertraf die Hauptkomponentenanalyse die *Treelet Transform* Methode in beiden Fällen hinsichtlich der erklärten Varianz um das doppelte. Der hohe Anteil erklärter Varianz bei der Hauptkomponentenanalyse erschließt sich dadurch, dass auf jeden extrahierten Faktor alle Originalvariablen laden, jeweils mit einer anderen Ladungsstärke. Auf die extrahierten Komponenten laden nur die Originalvariablen, welche auf Schnitthöhe des Dendrogrammes noch innerhalb einer korrelierten Gruppe lagen. Daher reduziert sich die Zahl der ladenden Variablen erheblich, aber ebenso der Anteil der erklärten Varianz.

Sowohl die Faktoren als auch die Komponenten zeigten zahlreiche Korrelationen mit anthropometrischen Parametern. Hinsichtlich der Tendenz bewegten sich die Korrelationskoeffizienten immer im gleichen Bereich. Lediglich hinsichtlich der Stärke unterschieden sich die Koeffizienten zwischen Komponenten und Anthropometrie im Vergleich zu Faktoren und Anthropometrie, wobei sowohl die Faktoren als auch die Komponenten in manchen Fällen stärker korreliert waren.

Insgesamt zeigt diese Untersuchung, dass die Methode des TT eine Alternative zu einer Hauptkomponentenanalyse bietet, wenn der Datensatz mehr Variablen als Beobachtungen beinhaltet. Die beobachteten Korrelationen müssen in anderen Studien umfangreich belegt werden. Sollte dies jedoch möglich sein, dann können sie für die Identifizierung von Hochrisikoprobanden hinsichtlich chronischer Erkrankungen genutzt werden.

References

1. (WHO), W.H.O. *Obesity: Preventing and managing the global epidemic. [Online]. Available: 2000*. 2013.
2. Ogden, C.L., et al., *Prevalence of childhood and adult obesity in the United States, 2011-2012.* JAMA, 2014. **311**(8): p. 806-14.
3. Ford, E.S., et al., *Healthy living is the best revenge: findings from the European Prospective Investigation Into Cancer and Nutrition-Potsdam study.* Arch Intern Med, 2009. **169**(15): p. 1355-62.
4. Eckel, R.H., et al., *Prevention Conference VII: Obesity, a worldwide epidemic related to heart disease and stroke: executive summary.* Circulation, 2004. **110**(18): p. 2968-75.
5. Kahn, S.E., R.L. Hull, and K.M. Utzschneider, *Mechanisms linking obesity to insulin resistance and type 2 diabetes.* Nature, 2006. **444**(7121): p. 840-6.
6. Vucenik, I. and J.P. Stains, *Obesity and cancer risk: evidence, mechanisms, and recommendations.* Ann N Y Acad Sci, 2012. **1271**: p. 37-43.
7. Mokdad, A.H., et al., *Actual causes of death in the United States, 2000.* JAMA, 2004. **291**(10): p. 1238-45.
8. Pischon, T., et al., *General and abdominal adiposity and risk of death in Europe.* N Engl J Med, 2008. **359**(20): p. 2105-20.
9. Musaad, S. and E.N. Haynes, *Biomarkers of obesity and subsequent cardiovascular events.* Epidemiol Rev, 2007. **29**: p. 98-114.
10. Aleksandrova, K., et al., *Circulating C-reactive protein concentrations and risks of colon and rectal cancer: a nested case-control study within the European Prospective Investigation into Cancer and Nutrition.* Am J Epidemiol, 2010. **172**(4): p. 407-18.
11. Herder, C., et al., *Immunological and cardiometabolic risk factors in the prediction of type 2 diabetes and coronary events: MONICA/KORA Augsburg case-cohort study.* PLoS One, 2011. **6**(6): p. e19852.
12. Montonen, J., et al., *Consumption of red meat and whole-grain bread in relation to biomarkers of obesity, inflammation, glucose metabolism and oxidative stress.* Eur J Nutr, 2013. **52**(1): p. 337-45.
13. Chorell, E., et al., *Physical fitness level is reflected by alterations in the human plasma metabolome.* Mol Biosyst, 2012. **8**(4): p. 1187-96.
14. Joyce, A.R. and B.O. Palsson, *The model organism as a system: integrating 'omics' data sets.* Nat Rev Mol Cell Biol, 2006. **7**(3): p. 198-210.
15. Lewis, G.D., A. Asnani, and R.E. Gerszten, *Application of metabolomics to cardiovascular biomarker and pathway discovery.* J Am Coll Cardiol, 2008. **52**(2): p. 117-23.
16. Nicholson, J.K., J.C. Lindon, and E. Holmes, *'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.* Xenobiotica, 1999. **29**(11): p. 1181-9.

17. Nicholson, J.K. and J.C. Lindon, *Systems biology: Metabonomics.* Nature, 2008. **455**(7216): p. 1054-6.
18. Smith, C.A., et al., *XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.* Anal Chem, 2006. **78**(3): p. 779-87.
19. Floegel, A., et al., *Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach.* Diabetes, 2013. **62**(2): p. 639-48.
20. Wang, Z., et al., *Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease.* Nature, 2011. **472**(7341): p. 57-63.
21. Tworoger, S. and S. Hankinson, *Use of biomarkers in epidemiologic studies: minimizing the influence of measurement error in the study design and analysis.* Cancer Causes Control, 2006. **17**: p. 889–899.
22. Lee, A., B. Nadler, and L. Wassermann, *Treelets - An adaptive multi-scale basis for sparse unordered data.* The Annals of Applied Statistics, 2008. **2**(2): p. 435-471.
23. Bohling, G., *Dimension reduction and Cluster analysis.* EECS, 2006. **833**.
24. Hoffmann, K., et al., *Application of a new statistical method to derive dietary patterns in nutritional epidemiology.* Am J Epidemiol, 2004. **159**(10): p. 935-44.
25. Xiang, J. *Grouping of correlated feature vectors using treelets*. 2010. 1 - 8.
26. Gorst-Rasmussen, A., et al., *Exploring dietary patterns by using the treelet transform.* Am J Epidemiol, 2011. **173**(10): p. 1097-104.
27. Lee, A. and B. Nadler, *Treelets | A Tool for Dimensionality Reduction and Multi-Scale Analysis of Unstructured Data.* Journal of Machine Learning, 2007. **2**: p. 259 - 266.
28. Fischer, K., et al., *Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons.* PLOS Medicine, 2014. **11**(2).
29. Rhee, E.P. and R.E. Gerszten, *Metabolomics and cardiovascular biomarker discovery.* Clin Chem, 2012. **58**(1): p. 139-47.
30. Li, M., et al., *Symbiotic gut microbes modulate human metabolic phenotypes.* Proc Natl Acad Sci U S A, 2008. **105**(6): p. 2117-22.
31. Calvani, R., et al., *Gut microbiome-derived metabolites characterize a peculiar obese urinary metabotype.* Int J Obes (Lond), 2010. **34**(6): p. 1095-8.
32. Martin, F.P., et al., *Metabolomic applications to decipher gut microbial metabolic influence in health and disease.* Front Physiol, 2012. **3**: p. 113.
33. Velagapudi, V.R., et al., *The gut microbiota modulates host energy and lipid metabolism in mice.* J Lipid Res, 2010. **51**(5): p. 1101-12.
34. Dahm, C.C., et al., *Adipose tissue fatty acid patterns and changes in anthropometry: a cohort study.* PLoS One, 2011. **6**(7): p. e22587.
35. Gopinath, B., et al., *Adiposity Adversely Influences Quality of Life Among Adolescents.* J Adolesc Health, 2013.
36. Diehr, P., et al., *Comparing years of healthy life, measured in 16 ways, for normal weight and overweight older adults.* J Obes, 2012. **2012**: p. 894894.
37. WHO, *http://www.who.int/mediacentre/factsheets/fs311/en/index.html.* 2013.
38. Correll, C.U., et al., *Findings of a U.S. national cardiometabolic screening program among 10,084 psychiatric outpatients.* Psychiatr Serv, 2010. **61**(9): p. 892-8.
39. Bradshaw, T. and H. Mairs, *Obesity and Serious Mental Ill Health: A Critical Review of the Literature.* Healthcare, 2014. **2**: p. 166-182.

40. Robert Koch Institute, R., *http://www.rki.de/DE/Content/Gesundheitsmonitoring/Studien/Degs/degs_w1/Symposium/degs_uebergewicht_adipositas.pdf?__blob=publicationFile.* 2012.
41. Destatis, *https: //www.genesis.destatis.de/genesis/online/data; jsessionid=DC9B7D363442 F4284E14EF6B1DCED86F.tomcat_GO_2_1?operation= abruftabelle Bearbeiten& levelindex =2&levelid=1364309364829&auswahloperation= abruf tabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel = werteabruf&selectionname=23631-0001&auswahltext=&werteabruf=starten.* 2013.
42. Kelly, T., et al., *Global burden of obesity in 2005 and projections to 2030.* Int J Obes (Lond), 2008. **32**(9): p. 1431-7.
43. Bouchard, C., *The biological predisposition to obesity: beyond the thrifty genotype scenario.* Int J Obes (Lond), 2007. **31**(9): p. 1337-9.
44. Farooqi, I.S. and S. O'Rahilly, *Genetic factors in human obesity.* Obes Rev, 2007. **8 Suppl 1**: p. 37-40.
45. Biondi, B., *Thyroid and obesity: an intriguing relationship.* J Clin Endocrinol Metab, 2010. **95**(8): p. 3614-7.
46. WHO, *Obesity: Preventing and managing the global epidemic; Part III Understanding how overweight and obesity develop.* 2000.
47. Hu, F.B., *Metabolic profiling of diabetes: from black-box epidemiology to systems epidemiology.* Clin Chem, 2011. **57**(9): p. 1224-6.
48. Shah, S.H., W.E. Kraus, and C.B. Newgard, *Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases: form and function.* Circulation, 2012. **126**(9): p. 1110-20.
49. Bühlmann, P. and S. van de Geer, *Statistics for High-Dimensional Data*. Methods, Theory and Applications Series: Springer Series in Statistics. 2011, Berlin Heidelberg: Springer. 574.
50. Madhulatha, T., *An overview on clustering methods.* IOSR Journal of Engineering, April 2012. **2**(4): p. 719 - 725.
51. Bingham, S. and E. Riboli, *Diet and cancer--the European Prospective Investigation into Cancer and Nutrition.* Nat Rev Cancer, 2004. **4**(3): p. 206-15.
52. Boeing, H., J. Wahrendorf, and N. Becker, *EPIC-Germany--A source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition.* Ann Nutr Metab, 1999. **43**(4): p. 195-204.
53. Oresic, M., et al., *Metabolome in schizophrenia and other psychotic disorders: a general population-based study.* Genome Med, 2011. **3**(3): p. 19.
54. WHO, *Physical Status: The Use and Interpretation of Anthropometry.* WHO Technical Report Series, 1995. **854**: p. Annex 2.
55. Durnin, J.V. and J. Womersley, *Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years.* Br J Nutr, 1974. **32**(1): p. 77-97.
56. Bergman, R.N., et al., *A better index of body adiposity.* Obesity (Silver Spring), 2011. **19**(5): p. 1083-9.
57. Wald, D., et al., *Automatic quantification of subcutaneous and visceral adipose tissue from whole-body magnetic resonance images suitable for large cohort studies.* J Magn Reson Imaging, 2012. **36**(6): p. 1421-1434.

58. Maukonen, J., et al., *PCR DGGE and RT-PCR DGGE show diversity and short-term temporal stability in the Clostridium coccoides-Eubacterium rectale group in the human intestinal microbiota.* FEMS Microbiol Ecol, 2006. **58**(3): p. 517-28.
59. Nubel, U., et al., *Sequence heterogeneities of genes encoding 16S rRNAs in Paenibacillus polymyxa detected by temperature gradient gel electrophoresis.* J Bacteriol, 1996. **178**(19): p. 5636-43.
60. Tourlomousis, P., et al., *PCR-denaturing gradient gel electrophoresis of complex microbial communities: a two-step approach to address the effect of gel-to-gel variation and allow valid comparisons across a large dataset.* Microb Ecol, 2010. **59**(4): p. 776-86.
61. Davenport, E.R., et al., *Seasonal variation in human gut microbiome composition.* PLoS One, 2014. **9**(3): p. e90731.
62. Wills, E.S., et al., *Fecal microbial composition of ulcerative colitis and Crohn's disease patients in remission and subsequent exacerbation.* PLoS One, 2014. **9**(3): p. e90981.
63. Gorst-Rasmussen, A., *tt: Treelet transform with Stata.* Stata Journal, 2012. **12**(1): p. 130-146.
64. Hoffmann, K., et al., *Comparison of two statistical approaches to predict all-cause mortality by dietary patterns in German elderly subjects.* Br J Nutr, 2005. **93**(5): p. 709-16.
65. Brown, J.D., *Choosing the Right Number of Components or Factors in PCA and EFA.* JALT Testing & Evaluation SIG Newsletter, 2009. **13**(2): p. 19-23.
66. Zhu, M. and A. Ghodsi, *Automatic dimensionality selection from the scree plot via the use of profile likelihood.* Computational Statistics & Data Analysis, 2006. **51**: p. 918 – 930.
67. Brown, J.D., *Choosing the Right Type of Rotation in PCA and EFA.* JALT Testing & Evaluation SIG Newsletter, 2009. **13**(3): p. 20-25.
68. Abdi, H., *Factor Rotations in Factor Analysis*, in *Encyclopedia of Social Sciences Research Methods*, M. Lewis-Beck, A. Bryman, and F. T., Editors. 2003, Thousand Oaks (CA): Sage.
69. Inc., S.I., *SAS/STAT® 9.22 User's Guide.* Cary, NC: SAS Institute Inc., 2010
70. Prokhorov, A., *Partial correlation coefficient.* Encyclopedia of Mathematics, 2011.
71. Mittelstrass, K., et al., *Discovery of sexual dimorphisms in metabolic and genetic biomarkers.* PLoS Genet, 2011. **7**(8): p. e1002215.
72. Yu, Z., et al., *Human serum metabolic profiles are age dependent.* Aging Cell, 2012. **11**(6): p. 960-7.
73. John Wiley & Sons, I., *Encyclopedia of Statistical Sciences*. 2006.
74. Mueller, S., et al., *Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study.* Appl Environ Microbiol, 2006. **72**(2): p. 1027-33.
75. Nwosu, F.C., et al., *Age-dependent fecal bacterial correlation to inflammatory bowel disease for newly diagnosed untreated children.* Gastroenterol Res Pract, 2013. **2013**: p. 302398.
76. Markle, J.G., et al., *Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity.* Science, 2013. **339**(6123): p. 1084-8.
77. Cohen, J., *A power primer.* Psychol Bull, 1992. **112**(1): p. 155-9.

78. Meinshausen, N. and P. Bühlmann, *Discussion of: Treelets-an adaptive multi-scale basis for sparse unordered data.* The Annals of Applied Statistics, 2008. **2**(2): p. 478 - 481.
79. Neamat-Allah, J., et al., *Validation of Anthropometric Indices of Adiposity against Whole-Body Magnetic Resonance Imaging - A Study within the German European Prospective Investigation into Cancer and Nutrition (EPIC) Cohorts.* PLoS One, 2014. **9**(3): p. e91586.
80. Vinayavekhin, N. and A. Saghatelian, *Untargeted metabolomics.* Curr Protoc Mol Biol, 2010. **Chapter 30**: p. Unit 30 1 1-24.
81. Breier, M., et al., *Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples.* PLoS One, 2014. **9**(2): p. e89728.
82. Müller, M.J. and E. Trautwein, *Gesundheit und Ernährung. Public Health Nutrition*. 2005, Stuttgart: Ulmer UTB. 304.
83. Oberbach, A., et al., *Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes.* J Proteome Res, 2011. **10**(10): p. 4769-88.
84. Newgard, C.B., et al., *A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance.* Cell Metab, 2009. **9**(4): p. 311-26.
85. Huffman, K.M., et al., *Relationships between circulating metabolic intermediates and insulin action in overweight to obese, inactive men and women.* Diabetes Care, 2009. **32**(9): p. 1678-83.
86. Cevallos-Cevallos, J.M. and J.I. Reyes-De-Corcuera, *Metabolomics in food science.* Adv Food Nutr Res, 2012. **67**: p. 1-24.
87. D'Alessandro, A., et al., *Clinical metabolomics: the next stage of clinical biochemistry.* Blood Transfus, 2012. **10 Suppl 2**: p. s19-24.
88. Vinayavekhin, N., E.A. Homan, and A. Saghatelian, *Exploring disease through metabolomics.* ACS Chem Biol, 2010. **5**(1): p. 91-103.
89. Dunn, W.B., et al., *A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols.* Int J Epidemiol, 2008. **37 Suppl 1**: p. i23-30.
90. Zanolin, M.E., et al., *Clustering of cardiovascular risk factors associated with the insulin resistance syndrome: assessment by principal component analysis in young hyperandrogenic women.* Diabetes Care, 2006. **29**(2): p. 372-8.
91. Tsukui, S., Y. Fukumura, and I. Kobayashi, *Decreased serum 1,5-anhydroglucitol in nondiabetic subjects with a family history of NIDDM.* Diabetes Care, 1996. **19**(9): p. 940-4.
92. Warensjo, E., et al., *Factor analysis of fatty acids in serum lipids as a measure of dietary fat quality in relation to the metabolic syndrome in men.* Am J Clin Nutr, 2006. **84**(2): p. 442-8.
93. Yang, J., et al., *Discrimination of Type 2 diabetic patients from healthy controls by using metabonomics method based on their serum fatty acid profiles.* J Chromatogr B Analyt Technol Biomed Life Sci, 2004. **813**(1-2): p. 53-8.
94. Dahm, C.C., et al., *Fatty acid patterns and risk of prostate cancer in a case-control study nested within the European Prospective Investigation into Cancer and Nutrition.* Am J Clin Nutr, 2012. **96**(6): p. 1354-61.

95. Schoenaker, D.A., et al., *Factor analysis is more appropriate to identify overall dietary patterns associated with diabetes when compared with Treelet transform analysis.* J Nutr, 2013. **143**(3): p. 392-8.
96. Schmidt, J.A., et al., *Baseline patterns of adipose tissue fatty acids and long-term risk of breast cancer: a case-cohort study in the Danish cohort Diet, Cancer and Health.* Eur J Clin Nutr, 2014.
97. Patti, G.J., O. Yanes, and G. Siuzdak, *Innovation: Metabolomics: the apogee of the omics trilogy.* Nat Rev Mol Cell Biol, 2012. **13**(4): p. 263-9.
98. Hankinson, S.E., et al., *Alcohol, height, and adiposity in relation to estrogen and prolactin levels in postmenopausal women.* J Natl Cancer Inst, 1995. **87**(17): p. 1297-302.
99. Jadhav, A.A. and A. Jain, *Elevated adenosine deaminase activity in overweight and obese Indian subjects.* Arch Physiol Biochem, 2012. **118**(1): p. 1-5.
100. Spencer, M.E., et al., *Serum levels of the immune activation marker neopterin change with age and gender and are modified by race, BMI, and percentage of body fat.* J Gerontol A Biol Sci Med Sci, 2010. **65**(8): p. 858-65.
101. Szymanska, E., et al., *Gender-dependent associations of metabolite profiles and body fat distribution in a healthy population with central obesity: towards metabolomics diagnostics.* OMICS, 2012. **16**(12): p. 652-67.
102. Elshorbagy, A.K., et al., *S-adenosylmethionine is associated with fat mass and truncal adiposity in older adults.* J Nutr, 2013. **143**(12): p. 1982-8.
103. Lustgarten, M.S., et al., *Serum glycine is associated with regional body fat and insulin resistance in functionally-limited older adults.* PLoS One, 2013. **8**(12): p. e84034.
104. Pennetti, V., et al., *Relation between obesity, insulinemia, and serum amino acid concentrations in a sample of Italian adults.* Clin Chem, 1982. **28**(11): p. 2219-24.
105. van Driel, L.M., et al., *Body mass index is an important determinant of methylation biomarkers in women of reproductive ages.* J Nutr, 2009. **139**(12): p. 2315-21.
106. Kroger, J., et al., *Erythrocyte membrane phospholipid fatty acids, desaturase activity, and dietary fatty acids in relation to risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study.* Am J Clin Nutr, 2011. **93**(1): p. 127-42.
107. Bennasar-Veny, M., et al., *Body adiposity index and cardiovascular health risk factors in Caucasians: a comparison with the body mass index and others.* PLoS One, 2013. **8**(5): p. e63999.
108. Sadeghi, M., et al., *Abdominal fat distribution and serum lipids in patients with and without coronary heart disease.* Arch Iran Med, 2013. **16**(3): p. 149-53.
109. Rocha, F.L., et al., *Correlation between indicators of abdominal obesity and serum lipids in the elderly.* Rev Assoc Med Bras, 2013. **59**(1): p. 48-55.
110. Jourdan, C., et al., *Body fat free mass is associated with the serum metabolite profile in a population-based study.* PLoS One, 2012. **7**(6): p. e40009.
111. Respondek, F., et al., *Short-Chain Fructo-Oligosaccharides Modulate Intestinal Microbiota and Metabolic Parameters of Humanized Gnotobiotic Diet Induced Obesity Mice.* Plos One, 2013. **8**(8).
112. Moschen, A.R., V. Wieser, and H. Tilg, *Dietary Factors: Major Regulators of the Gut's Microbiota.* Gut Liver, 2012. **6**(4): p. 411-6.

113. Hoffmann, C., et al., *Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents.* Plos One, 2013. **8**(6).
114. Foerster, J., et al., *The influence of whole grain products and red meat on intestinal microbiota composition in normal weight adults: a randomized crossover intervention trial*. 2014: submitted.
115. Vrieze, A., et al., *The environment within: how gut microbiota may influence metabolism and body composition.* Diabetologia, 2010. **53**(4): p. 606-13.
116. Matthies, A., et al., *Daidzein and genistein are converted to equol and 5-hydroxy-equol by human intestinal Slackia isoflavoniconvertens in gnotobiotic rats.* J Nutr, 2012. **142**(1): p. 40-6.
117. Ursell, L.K., et al., *The intestinal metabolome: an intersection between microbiota and host.* Gastroenterology, 2014. **146**(6): p. 1470-6.
118. Russell, W.R., et al., *Colonic bacterial metabolites and human health.* Curr Opin Microbiol, 2013. **16**(3): p. 246-54.
119. Khanna, S. and P.K. Tosh, *A clinician's primer on the role of the microbiome in human health and disease.* Mayo Clin Proc, 2014. **89**(1): p. 107-14.
120. O'Connor, E.M., *The role of gut microbiota in nutritional status.* Current Opinion in Clinical Nutrition and Metabolic Care, 2013. **16**(5): p. 509-516.
121. Conterno, L., et al., *Obesity and the gut microbiota: does up-regulating colonic fermentation protect against obesity and metabolic disease?* Genes Nutr, 2011. **6**(3): p. 241-60.
122. Sanz, Y., A. Santacruz, and P. Gauffin, *Gut microbiota in obesity and metabolic disorders.* Proc Nutr Soc, 2010. **69**(3): p. 434-41.
123. Macfarlane, G.T. and S. Macfarlane, *Bacteria, colonic fermentation, and gastrointestinal health.* J AOAC Int, 2012. **95**(1): p. 50-60.
124. Nieuwdorp, M., et al., *Role of the microbiome in energy regulation and metabolism.* Gastroenterology, 2014. **146**(6): p. 1525-33.
125. Tabbaa, M., et al., *Docosahexaenoic Acid, Inflammation, and Bacterial Dysbiosis in Relation to Periodontal Disease, Inflammatory Bowel Disease, and the Metabolic Syndrome.* Nutrients, 2013. **5**(8): p. 3299-3310.
126. Foerster, J., *Intestinal Microbiota and Obesity*, in *Charite Neuroscience Newsletter*. December 2012. p. 10.
127. Vaarala, O., *Human intestinal microbiota and type 1 diabetes.* Curr Diab Rep, 2013. **13**(5): p. 601-7.
128. Macfarlane, G. and G. Gibson, *Carbohydrate fermentation, energy transduction and gas metabolism in the human large intestine*, in *Gastrointestinal microbiology*, R. Mackie and B. White, Editors. 1997, Chapman and Hall: New York. p. 269–318.
129. Lattimer, J.M. and M.D. Haub, *Effects of dietary fiber and its components on metabolic health.* Nutrients, 2010. **2**(12): p. 1266-89.
130. Jumpertz, R., et al., *Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans.* Am J Clin Nutr, 2011. **94**(1): p. 58-65.
131. Cox, L.M. and M.J. Blaser, *Pathways in microbe-induced obesity.* Cell Metab, 2013. **17**(6): p. 883-94.
132. Ling, Z., et al., *Analysis of oral microbiota in children with dental caries by PCR-DGGE and barcoded pyrosequencing.* Microb Ecol, 2010. **60**(3): p. 677-90.

133. Zwielehner, J., et al., *Changes in human fecal microbiota due to chemotherapy analyzed by TaqMan-PCR, 454 sequencing and PCR-DGGE fingerprinting.* PLoS One, 2011. **6**(12): p. e28654.
134. Fraher, M.H., P.W. O'Toole, and E.M. Quigley, *Techniques used to characterize the gut microbiota: a guide for the clinician.* Nat Rev Gastroenterol Hepatol, 2012. **9**(6): p. 312-22.
135. Ahn, J., et al., *Human gut microbiome and risk for colorectal cancer.* J Natl Cancer Inst, 2013. **105**(24): p. 1907-11.
136. Jakobsson, H.E., et al., *Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by Caesarean section.* Gut, 2014. **63**(4): p. 559-66.
137. van Baarlen, P., M. Kleerebezem, and J.M. Wells, *Omics approaches to study host-microbiota interactions.* Curr Opin Microbiol, 2013. **16**(3): p. 270-7.
138. Molla, M.T., C. Centers for Disease, and Prevention, *Expected years of life free of chronic condition-induced activity limitations - United States, 1999-2008.* MMWR Surveill Summ, 2013. **62 Suppl 3**: p. 87-92.
139. Holick, M.F., *Vitamin D: importance in the prevention of cancers, type 1 diabetes, heart disease, and osteoporosis.* Am J Clin Nutr, 2004. **79**(3): p. 362-71.

Supplement

Descriptive statistics of small polar metabolites detected in serum

TABLE S 1: Description of amino acids and amino acid derivates

Small polar metabolite (variable number)	Mean	Mini-mum	Maxi-mum	Variance	ln variance
Alanine (20)	405.33	11.59	730.20	11670.11	0.12
Alanine, phenyl-, dl (103)	550.64	220.96	1232.33	31126.58	0.10
Alpha-ketoglutaric acid (540)	37.08	0.00	209.53	1645.81	3.54
Aminomalonic acid (41)	445.81	158.12	1627.16	46423.12	0.17
Aspartic acid (18)	53.09	20.03	126.21	388.47	0.13
Glutamic acid (2)	343.22	161.70	790.73	8983.42	0.07
Glutamine (361)	88.90	0.00	513.72	8358.90	2.41
Glycine (24)	203.07	99.07	385.00	2580.25	0.06
Glycine (567)	4.18	0.00	32.13	33.67	1.06
Hippuric acid (462)	22.01	0.00	141.83	650.34	1.95
Isoleucine (17)	106.27	0.00	283.76	1647.56	0.23
Leucine (15)	240.05	107.39	584.97	5262.55	0.08
L-Proline, 4-hydroxy-, trans- (137)	57.34	6.61	320.88	2071.51	0.41
l-Threonine (184)	1532.75	501.93	6476.42	353678.47	0.11
L-Tryptophan (256)	86.05	0.00	297.01	3453.97	1.60
l-Valine (1211)	813.90	0.00	4811.06	270221.41	1.97
Methionine (1)	171.73	74.72	323.09	1315.49	0.05
Ornithine (10)	603.68	0.00	2137.22	194119.26	1.02
Phenylalanine (14)	362.60	230.70	509.79	2804.74	0.02
Proline (19)	388.77	132.87	716.28	12236.99	0.09
Proline* (88)	97.28	12.44	596.35	4853.36	0.33
Pyroglutamic acid (98)	3899.82	1978.09	7730.26	843199.00	0.05
Serine (4)	254.42	111.71	406.31	2251.53	0.04
Serine* (176)	1341.16	373.89	12195.22	1473211.23	0.22
Threonine (9)	270.53	128.87	460.25	3985.81	0.06
Tryptophan (287)	536.05	82.06	1758.02	47563.69	0.19
Tyrosine (13)	529.25	0.00	11988.94	665656.96	0.71
Tyrosine (64)	708.29	177.06	2222.33	70541.35	0.13
Valine (23)	269.70	139.70	414.26	2931.00	0.04
α-Alanine (175)	74.87	0.00	438.77	3158.88	0.35

TABLE S 2: Description of fatty acids and fatty acid derivates

Small polar metabolite (variable number)	Mean	Mini-mum	Maxi-mum	Variance	ln variance
2-Butenedioic acid (107)	20.95	7.53	88.62	76.04	0.12
2-Ethylhexanoic acid (527)	165.84	0.00	1104.30	31459.37	5.71
3-Hydroxycaproic acid (409)	23.65	0.00	157.15	284.33	1.45
4-Hydroxyvaleric acid (590)	7.74	0.00	125.41	150.38	1.72
4-Methyl-2-oxovaleric acid (192)	457.21	0.00	1340.37	62869.14	0.69
9-Tetradecenoic acid (360)	35.11	0.00	167.05	786.71	1.28
Arachidonic acid* (1192)	93.23	0.00	354.98	3110.62	0.71
Arachidonic acid* (1346)	40.85	0.00	126.36	951.46	3.28
Arachidonic acid (3)	84.94	29.76	153.61	654.93	0.09
Benzeneacetic acid, 4-hydroxy (209)	38.59	0.00	99.41	247.47	0.27
Decanoic acid (33)	89.27	15.08	796.52	7414.90	0.39
Heptanoic acid (438)	65.44	0.00	217.46	2924.06	2.31
Hexadecanoic acid, 2,3-bishydroxy (305)	149.46	53.76	453.64	4377.37	0.17
Hexadecanoic acid, 3,7,11,15-tetramethyl (300)	7.92	0.00	42.48	31.08	0.45
Hexanoic acid (599)	395.91	0.00	35990.57	5780431.73	8.35
Lauric acid (179)	146.90	25.47	1109.05	20127.20	0.39
Linoleic acid* (1185)	49.89	0.00	164.42	546.41	0.47
Linoleic acid* (134)	89.54	0.00	270.72	2003.23	0.35
Linoleic acid* (157)	61.03	0.00	128.69	463.67	0.26
Linoleic acid (7)	293.54	88.30	943.89	8690.11	0.09
Linolenic acid (159)	195.33	0.00	2161.88	33695.42	0.50
Myristic acid (72)	2227.92	63.69	64871.78	68491661.05	1.30
Nonanoic acid (89)	198.81	50.86	601.05	11159.14	0.28
Octanoic acid (263)	138.54	33.20	651.11	6076.77	0.26
Oleic acid (12)	577.69	199.67	1928.29	52697.66	0.16
Palmitelaidic acid (292)	678.14	0.00	3297.83	435806.21	2.07
Palmitic acid (8)	417.99	176.37	1042.32	12687.96	0.07
Pentanoic acid, 3-methyl-2-hydroxy (534)	13.43	0.00	133.14	199.49	1.94
Stearic acid* (112)	12.76	6.93	24.41	8.06	0.04
Stearic acid* (163)	11.06	0.00	27.96	15.54	0.14
Stearic acid* (270)	72.24	0.00	221.78	896.35	0.22
Stearic acid* (530)	4.67	0.00	16.58	13.22	0.90
Stearic acid (6)	249.23	115.53	551.82	3899.92	0.06

TABLE S 3: Description of carboxylic acid compounds

Small polar metabolite (variable number)	Mean	Mini-mum	Maxi mum	Variance	ln variance
1,3-Benzenedicarboxylic acid (397)	3.14	0.00	6.55	1.93	0.25
1H-Indole-3-acetic acid (63)	53.99	7.95	549.65	3169.56	0.44
1-Propene-1,2,3-tricarboxylic acid (259)	33.73	0.00	852.68	6593.25	0.58
2- Aminobutyric acid (340)	238.62	0.00	803.45	14322.73	1.43
2,3,4-Trihydroxybutyric acid (47)	1716.92	344.13	3337.58	285173.87	0.12
2,3-Dihydroxybutanoic acid (117)	54.48	11.97	175.77	610.67	0.16
2,3-Dihydroxybutanoic acid* (1190)	93.08	0.00	459.10	2885.65	0.64
2,5-Furandicarboxylic acid (556)	8.13	0.00	87.74	133.41	1.57
2-Ethyl-3-hydroxypropionic acid (284)	72.56	0.00	546.18	2257.40	0.57
2-Furancarboxylic acid, 5-hydroxy (600)	6.21	0.00	83.37	113.31	1.53
2-Hydroxybutyric acid (294)	1804.42	545.77	5598.50	849233.20	0.20
3-Hydroxybutyric acid (21)	83.98	21.68	907.08	7299.31	0.39
3-Hydroxyphenylpropionic acid (581)	11.20	0.00	116.89	335.98	2.16
3-Phenylpropionic acid (2103)	13.52	0.00	120.82	343.45	2.11
Adipic acid (288)	10.05	0.00	85.09	70.65	0.43
Benzeneacetic acid (492)	18.85	0.00	182.64	380.64	1.71
Butanedioic acid, methylene- (559)	12.45	0.00	85.89	220.58	1.98
Butanoic acid, 2,4-bishydroxy (78)	43.99	12.30	160.45	490.26	0.21
Butanoic acid, 3-methyl-2-hydroxy (504)	280.14	0.00	2064.06	120754.01	5.63
Butanoic acid, 3-methyl-3-hydroxy (421)	42.86	0.00	210.35	1093.43	2.25
Butanoic acid, 4-hydroxy (614)	7.30	0.00	37.10	55.84	1.71
Citric acid (16)	269.17	100.59	768.76	7798.58	0.10
Glycolic acid (365)	184.60	0.00	1148.33	35512.41	6.44
Lactic acid (26)	1187.33	418.08	2260.18	125370.86	0.09
Lactic acid* (471)	2033.82	0.00	14844.02	3852422.80	9.55
Malic acid (95)	88.94	31.68	481.72	2227.17	0.15
Propanedioic acid (1264)	4.91	0.00	34.01	37.30	1.37
Propanoic acid, 2-oxo-3-hydroxy (1206)	659.20	0.00	5740.01	670494.59	2.34
β-Amino isobutyric acid (282)	6.36	0.00	25.25	21.90	0.44

TABLE S 4: Description of carbohydrates and carbohydrate derivates

Small polar metabolite (variable number)	Mean	Mini-mum	Maxi mum	Variance	ln variance
Arabinofuranose (144)	11.71	3.14	56.54	43.30	0.17
D-fructose (75)	21785.15	2148.84	3077758.52	44344135165.00	0.44
D-fructose* (111)	127.54	41.20	629.79	5062.49	0.20
D-fructose* (1182)	3317.86	0.00	9829.87	1351207.49	0.66
D-galactose (337)	10.51	0.00	35.63	38.49	0.46
D-galactose* (1200)	16556.80	0.00	3275712.81	47471561406.00	2.52
D-glucitol, 6-deoxy (417)	14.04	0.00	171.11	480.59	1.05
Erythrose (480)	21.90	0.00	326.29	1606.10	1.89
Erythrose* (1428)	9245.76	0.00	238446.18	367022415.00	22.38
Gluconic acid, 2-oxo (553)	202.50	0.00	2064.78	78534.51	7.16
Glucopyranose (167)	5593.13	0.00	13750.89	6262915.96	0.71
Glucopyranose* (206)	7270.33	766.84	28144.67	13031463.88	0.32
Glucose (1339)	32871.26	0.00	1675486.98	12642384650.00	1.70
Glucose* (1394)	4285.90	0.00	81712.68	60992229.31	4.51
Glucose* (1418)	7636.69	0.00	45883.12	63425714.01	11.45
Glucose* (1424)	13549.73	0.00	111926.51	343263382.00	19.30
Glucose* (1430)	4765.22	0.00	37335.31	57952268.17	19.46
Glucose* (252)	13597.94	0.00	79838.07	77460517.10	1.06
Glyceric acid (85)	363.72	144.42	952.61	18195.18	0.12
Glycoside, α-methyl (1232)	109.28	0.00	3251.69	76811.27	3.43
Lyxose (514)	10.02	0.00	243.87	992.72	1.22
Pentitol, 1-desoxy (241)	12.32	0.00	50.79	63.39	0.35
Ribonic acid (1189)	41.48	0.00	182.45	842.09	0.64
Ribonic acid* (118)	233.56	16.00	3169.07	152014.73	1.22
Xylitol (102)	24.64	5.77	254.71	301.41	0.13
Xylitol* (135)	226.94	60.63	669.84	8816.35	0.13

TABLE S 5: Description of other small polar metabolites

Small polar metabolite (variable number)	Mean	Mini-mum	Maxi mum	Variance	ln variance
1-Dodecanol (374)	6.25	0.00	28.84	12.87	0.42
3,7-Dioxa-2,8-disilanonane (1222)	25.20	0.00	91.89	385.27	1.63
3,8-Dioxa-2,9-disiladecane (123)	296.30	90.09	3466.54	54577.00	0.14
3-Methyl-1,3-bisbutane (593)	13.50	0.00	146.51	362.23	2.51

Butane, 1,2,3-trishydroxy (1187)	32.83	0.00	160.53	480.36	0.41
Cholesterol (25)	3739.28	1677.55	6225.65	486773.85	0.04
Creatinine (54)	708.17	91.18	29463.70	4366472.27	0.38
Ethanolamine (79)	138.98	66.29	257.25	1575.83	0.08
Glycerol (149)	3387.00	1059.40	7428.45	1705992.42	0.15
Hypoxanthine (390)	18.24	0.00	79.30	206.52	1.22
Myo-Inositol (32)	2367.12	1344.75	5327.43	360148.50	0.05
Phosphoric acid monomethyl ester (264)	365.96	122.15	887.96	15854.54	0.12
tert-Butylpentamethyldisiloxane (467)	78.45	0.00	531.13	8825.84	3.11
Urea (343)	6676.04	0.00	108305.77	144936275.00	1.22
Urea* (1395)	13097.04	0.00	91043.56	377075113.00	12.04
Urea* (1459)	2066.55	0.00	61576.58	65164788.98	11.27

Descriptive statistics of lipids detected in serum

TABLE S 6: Description of Cholesteryl esters

Lipid (variable number)	Mean	Minimum	Maximum	Variance
ChoE(14:0) (460)	1.56	0.50	2.51	0.13
ChoE(16:0) (176)	8.72	4.54	14.28	2.86
ChoE(16:1) (339)	2.65	0.63	10.48	2.17
ChoE(18:1) (50)	48.62	27.65	80.83	92.09
ChoE(18:2) (17)	136.36	56.65	202.23	649.64
ChoE(18:3) (177)	5.51	0.70	18.25	8.25
ChoE(20:3) (364)	3.19	1.17	5.87	0.64
ChoE(20:4) (43)	41.26	16.87	83.11	148.44
ChoE(20:5) (150)	7.93	0.86	28.37	25.99

TABLE S 7: Description of Ceramides

Lipid (variable number)	Mean	Minimum	Maximum	Variance
Cer(d18:1/16:0) (707)	0.44	0.12	0.92	0.02
Cer(d18:1/17:0) (292)	2.26	1.78	3.06	0.07
Cer(d18:1/16:1) (1311)	3.47	1.89	6.07	0.35
Cer(d18:1/22:1) (965)	0.25	0.11	0.62	0.01
Cer(d18:1/24:2) (665)	0.37	0.16	0.84	0.01
Cer(d18:1/22:6) (106)	11.53	7.55	16.25	2.98
Cer(d18:1/22:6) (550)	0.85	0.42	1.54	0.04

TABLE S 8: Description of Di- and Triglycerides

Lipid (variable number)	Mean	Minimum	Maximum	Variance
DG(28:2) (426)	0.51	0.17	1.09	0.03
DG(36:3) (1167)	0.08	0.02	0.27	0.00
DG(33:5) (462)	0.30	0.21	0.45	0.00
DG(40:7) (997)	0.11	0.04	0.45	0.00
TG(16:0/18:1/16:0) (501)	0.85	0.61	1.29	0.02
TG(42:0) (536)	1.72	0.08	29.45	9.46
TG(44:0) (1086)	0.56	0.01	6.38	0.64
TG(44:0) (350)	4.06	0.17	38.99	26.48
TG(45:0) (528)	0.45	0.07	2.88	0.24
TG(46:0) (1179)	1.93	0.08	10.99	3.05
TG(46:0) (1265)	6.50	0.23	39.72	54.94
TG(47:0) (455)	0.88	0.11	5.23	0.87
TG(47:0) (747)	0.32	0.04	1.62	0.09
TG(48:0) (147)	13.18	0.57	67.72	144.55
TG(49:0) (530)	0.85	0.08	5.42	0.92
TG(49:0) (830)	0.97	0.11	3.91	0.57
TG(49:0)* (1045)	0.86	0.06	6.03	0.84
TG(50:0) (1184)	0.65	0.15	2.27	0.21
TG(50:0) (1300)	12.54	1.13	30.72	30.89
TG(50:0) (195)	9.56	0.28	62.76	116.46
TG(52:0) (137)	2.87	1.74	4.08	0.20
TG(52:0) (162)	6.51	0.30	20.61	18.35
TG(52:0) (472)	2.01	0.07	19.01	10.56
TG(52:0) (591)	0.20	0.05	0.41	0.01
TG(53:0) (1014)	0.07	0.01	0.37	0.00
TG(53:0) (1172)	0.19	0.01	0.58	0.02
TG(54:0) (842)	0.77	0.04	5.52	0.87
TG(54:0) (886)	0.37	0.24	0.48	0.00
TG(54:0) (951)	0.14	0.01	1.23	0.03
TG(54:0) (953)	0.16	0.02	1.16	0.04
TG(42:1) (803)	0.77	0.01	22.50	3.37
TG(44:1) (437)	3.65	0.06	41.94	26.37
TG(44:1) (732)	0.14	0.05	0.34	0.00
TG(45:1) (493)	0.51	0.04	4.05	0.32
TG(46:1) (184)	11.57	0.45	70.44	114.06
TG(46:1) (717)	0.20	0.08	0.49	0.01
TG(46:1) (821)	0.76	0.02	5.42	0.68
TG(47:1) (320)	1.88	0.24	8.75	2.80
TG(48:1) (1206)	3.35	0.31	13.00	4.74
TG(48:1) (48)	38.06	2.58	128.14	570.52
TG(49:1) (210)	5.34	0.64	22.71	17.03
TG(49:1) (837)	0.71	0.09	3.26	0.33
TG(50:1) (1022)	10.39	1.70	29.94	17.44

TG(50:1) (23)	71.20	6.56	188.99	1167.19
TG(51:1) (268)	3.46	0.35	14.00	7.38
TG(51:1) (313)	1.70	0.20	5.01	0.81
TG(52:1) (1017)	24.50	7.12	42.11	36.32
TG(52:1) (51)	33.26	1.74	106.03	541.64
TG(53:1) (1288)	1.03	0.13	3.31	0.45
TG(53:1) (410)	0.48	0.00	1.36	0.01
TG(53:1) (725)	0.14	0.02	0.79	0.01
TG(53:1) (880)	0.37	0.03	2.35	0.18
TG(54:1) (1274)	2.01	0.40	3.73	0.77
TG(54:1) (293)	3.41	0.12	23.81	16.01
TG(55:1) (776)	0.13	0.02	0.66	0.01
TG(55:1) (870)	0.42	0.06	0.70	0.02
TG(56:1) (1192)	0.25	0.02	2.64	0.12
TG(56:1) (915)	0.16	0.01	1.48	0.04
TG(56:1) (986)	0.08	0.01	0.60	0.01
TG(57:1) (1101)	0.12	0.01	0.47	0.01
TG(42:2) (827)	0.22	0.01	6.95	0.27
TG(44:2) (807)	0.83	0.02	12.57	1.78
TG(46:2) (1095)	0.12	0.04	0.26	0.00
TG(46:2) (346)	4.53	0.13	34.03	23.91
TG(47:2) (492)	0.72	0.10	3.72	0.41
TG(47:2) (967)	0.11	0.04	0.26	0.00
TG(48:2) (83)	18.91	2.05	76.57	159.52
TG(49:2) (223)	3.86	0.52	14.76	8.58
TG(49:2) (537)	0.32	0.14	0.64	0.01
TG(50:2) (1330)	0.10	0.06	0.22	0.00
TG(50:2) (14)	58.15	10.37	198.30	656.62
TG(50:2) (335)	0.42	0.08	0.91	0.03
TG(51:2) (116)	8.71	1.25	29.78	23.20
TG(51:2) (677)	0.47	0.05	1.89	0.13
TG(52:2) (5)	141.13	37.78	259.39	1679.10
TG(52:2) (542)	0.18	0.05	0.37	0.00
TG(53:2) (1138)	0.56	0.11	1.69	0.12
TG(54:2) (33)	34.33	6.34	102.22	373.28
TG(54:2) (365)	0.64	0.14	2.31	0.14
TG(55:2) (359)	0.42	0.10	0.69	0.02
TG(56:2) (1052)	0.45	0.09	2.29	0.13
TG(56:2) (474)	1.13	0.10	11.66	2.22
TG(56:2) (546)	0.23	0.06	0.86	0.01
TG(57:2) (885)	0.34	0.08	0.62	0.02
TG(58:2) (894)	0.12	0.01	1.52	0.03
TG(46:3) (695)	0.15	0.06	0.40	0.00
TG(46:3) (826)	0.62	0.03	7.76	0.73
TG(46:3) (875)	0.61	0.19	1.03	0.03
TG(48:3) (232)	6.06	0.41	36.33	25.06
TG(49:3) (432)	0.93	0.15	3.28	0.34

TG(50:3) (30)	31.93	8.75	66.61	168.96
TG(51:3) (164)	5.74	0.90	17.68	9.04
TG(52:3) (1113)	8.98	2.64	22.06	12.03
TG(52:3) (6)	121.57	48.79	233.57	1140.61
TG(52:3) (666)	0.22	0.04	0.44	0.01
TG(53:3) (1058)	0.32	0.07	1.19	0.03
TG(53:3) (1075)	0.17	0.00	0.69	0.01
TG(53:3) (1203)	2.70	0.60	6.96	1.93
TG(54:3) (10)	64.97	15.87	154.55	634.48
TG(54:3) (933)	0.14	0.02	0.40	0.01
TG(55:3) (384)	0.40	0.12	0.67	0.01
TG(55:3) (662)	0.25	0.07	0.54	0.01
TG(56:3) (258)	2.03	0.38	9.10	2.24
TG(58:3) (918)	0.20	0.02	3.36	0.12
TG(46:4) (57)	8.11	3.96	14.68	4.01
TG(47:4) (100)	3.15	1.25	7.02	0.87
TG(48:4) (1070)	0.39	0.14	1.13	0.03
TG(48:4) (84)	4.60	1.91	11.46	1.88
TG(48:4) (872)	1.34	0.49	2.54	0.13
TG(49:4) (599)	0.18	0.07	0.39	0.00
TG(51:4) (1207)	0.66	0.09	2.88	0.16
TG(51:4) (1247)	0.21	0.09	0.40	0.00
TG(51:4) (419)	1.09	0.24	3.95	0.34
TG(52:4) (21)	46.77	13.27	120.47	350.44
TG(52:4) (482)	0.27	0.08	0.43	0.01
TG(53:4) (269)	1.62	0.36	4.93	0.64
TG(54:4) (1269)	0.36	0.12	1.19	0.03
TG(54:4) (22)	44.62	12.49	150.68	433.30
TG(55:4) (406)	0.39	0.13	0.72	0.01
TG(55:4) (755)	0.16	0.03	0.39	0.00
TG(56:4) (1227)	0.43	0.12	0.79	0.02
TG(56:4) (261)	1.75	0.40	8.53	1.59
TG(56:4)* (447)	0.42	0.03	1.20	0.05
TG(57:4) (1353)	0.13	0.01	0.78	0.01
TG(57:4) (964)	0.22	0.08	0.42	0.01
TG(58:4) (969)	0.16	0.03	2.57	0.06
TG(48:5) (1173)	16.76	7.84	31.47	20.47
TG(49:5) (898)	0.57	0.23	2.25	0.05
TG(50:5) (766)	0.10	0.04	1.15	0.01
TG(50:5)* (633)	0.51	0.00	2.55	0.17
TG(51:5) (1079)	0.11	0.02	0.54	0.01
TG(51:5) (287)	2.19	0.77	3.44	0.38
TG(52:5) (138)	9.56	1.62	32.96	40.96
TG(53:5) (576)	0.42	0.09	1.35	0.05
TG(54:5) (54)	20.29	4.27	116.61	188.30
TG(54:5)* (800)	14.67	1.99	42.32	41.84
TG(55:5) (1008)	0.21	0.06	0.62	0.01

TG(56:5) (109)	3.92	0.87	11.98	3.42
TG(56:5) (1202)	2.62	0.46	6.14	0.89
TG(56:5) (828)	1.32	0.44	5.97	0.60
TG(58:5) (711)	0.12	0.04	0.32	0.00
TG(58:5) (958)	0.13	0.04	1.03	0.01
TG(59:5) (1063)	0.19	0.00	2.09	0.06
TG(59:5) (548)	0.17	0.06	0.46	0.00
TG(51:6) (640)	0.18	0.06	0.54	0.00
TG(52:6) (423)	0.91	0.13	6.62	0.52
TG(53:6) (136)	3.18	1.21	4.49	0.20
TG(54:6) (1020)	8.62	0.85	49.54	70.42
TG(54:6)* (97)	8.19	1.56	30.46	16.32
TG(55:6) (635)	0.21	0.04	1.09	0.02
TG(56:6) (244)	1.36	0.27	4.34	0.53
TG(56:6) (53)	11.18	2.41	27.93	16.89
TG(58:6) (1057)	0.31	0.14	1.34	0.02
TG(58:6) (1069)	0.18	0.07	0.38	0.00
TG(58:6) (310)	0.45	0.12	1.06	0.03
TG(59:6) (1067)	0.20	0.04	0.38	0.00
TG(51:7) (389)	0.49	0.21	0.93	0.02
TG(52:7) (785)	0.20	0.02	1.10	0.03
TG(53:7) (1110)	0.12	0.05	0.28	0.00
TG(53:7) (1186)	0.23	0.08	0.43	0.00
TG(53:7) (754)	0.10	0.04	0.20	0.00
TG(54:7) (1029)	1.06	0.00	33.28	8.12
TG(54:7) (216)	2.46	0.30	13.11	4.51
TG(55:7) (956)	0.20	0.00	0.90	0.03
TG(56:7) (808)	5.88	1.41	20.62	9.19
TG(56:7)* (1347)	14.01	1.87	45.72	65.99
TG(58:7) (1031)	0.88	0.32	1.91	0.09
TG(59:7) (1190)	0.22	0.04	0.53	0.01
TG(53:8) (205)	1.38	0.42	2.68	0.16
TG(56:8)* (1220)	5.58	0.48	21.63	18.13
TG(54:9) (173)	4.68	0.70	15.95	7.91
TG(56:9) (861)	0.41	0.05	1.57	0.09
TG(58:10)* (1250)	0.19	0.03	0.42	0.00
TG(58:10)* (838)	0.30	0.03	1.35	0.05

TABLE S 9: Description of Phosphatidylcholines and Lysophosphatidylcholines

Lipid (variable number)	Mean	Minimum	Maximum	Variance
PC(30:0) (1042)	1.08	0.61	2.52	0.10
PC(32:0) (69)	8.36	5.02	11.64	1.83
PC(33:0) (912)	0.65	0.27	1.48	0.07
PC(40:0) (733)	0.14	0.06	0.23	0.00
PC(30:1) (1019)	16.92	10.26	28.40	8.53
PC(31:1)+PE(34:1) (174)	1.86	0.92	3.32	0.20
PC(32:1) (40)	16.58	3.56	52.61	67.18
PC(34:1) (431)	0.57	0.23	1.51	0.05
PC(32:2) (148)	3.55	1.11	9.73	2.48
PC(32:2) (281)	1.10	0.49	3.21	0.10
PC(33:2)+PE(36:2) (110)	3.93	1.81	15.02	1.98
PC(34:2) (0)	192.13	137.27	246.76	452.09
PC(34:2) (1280)	0.19	0.08	0.88	0.01
PC(35:2) (78)	6.90	2.38	11.95	3.50
PC(37:2) (230)	1.32	0.54	3.11	0.15
PC(37:2) (308)	0.83	0.28	1.63	0.06
PC(38:2) (111)	4.43	2.28	8.18	1.09
PC(38:2) (1123)	1.42	0.64	2.88	0.17
PC(38:2e) (817)	4.14	1.74	10.85	1.29
PC(30:3) (342)	0.39	0.13	0.85	0.02
PC(32:3) (175)	2.52	1.10	7.19	0.47
PC(34:3) (108)	6.99	1.74	33.10	13.79
PC(34:3) (144)	3.31	0.80	14.20	2.88
PC(34:3) (156)	2.60	1.08	5.49	0.55
PC(34:3) (89)	4.26	1.82	9.69	2.59
PC(34:3e)+PE(37:3e) (169)	2.26	1.46	3.34	0.11
PC(34:3e)+PE(37:3e) (55)	7.29	2.48	12.63	4.10
PC(36:3) (44)	11.72	6.20	22.58	7.30
PC(36:3) (8)	68.63	33.17	110.18	222.64
PC(36:3) (93)	4.69	2.44	7.11	0.50
PC(38:3) (15)	39.59	14.28	80.79	112.89
PC(38:3e) (1044)	1.10	0.36	2.27	0.12
PC(38:3e) (1128)	0.51	0.21	0.87	0.02
PC(40:3e) (415)	0.42	0.17	0.77	0.01
PC(40:3e) (564)	0.27	0.17	0.39	0.00
PC(32:4) (649)	0.16	0.07	0.43	0.00
PC(34:4) (1156)	0.10	0.02	0.41	0.00
PC(34:4) (286)	0.45	0.30	1.16	0.01
PC(36:4) (1)	147.47	76.52	205.91	443.27
PC(36:4) (1036)	1.06	0.71	1.43	0.02
PC(36:4) (36)	17.64	4.96	53.98	53.76
PC(36:4e) (1065)	0.47	0.20	1.24	0.03
PC(36:4e) (262)	1.03	0.43	2.65	0.12
PC(37:4)/PE(40:4) (73)	6.76	2.88	12.19	3.24
PC(38:4) (1025)	5.13	1.62	12.35	3.32
PC(38:4) (4)	96.09	38.11	148.10	393.85

PC(38:4) (80)	7.63	2.08	18.29	6.84
PC(40:4) (167)	2.79	0.58	8.89	1.35
PC(40:4e) (1185)	0.25	0.10	0.46	0.01
PC(40:4e) (297)	0.93	0.33	1.74	0.07
PC(40:4e) (400)	0.42	0.21	0.64	0.01
PC(32:5) (161)	1.19	0.49	2.65	0.16
PC(32:5) (361)	0.91	0.24	1.67	0.08
PC(36:5) (1018)	14.58	3.44	45.03	48.86
PC(36:5) (1127)	0.64	0.32	1.48	0.03
PC(36:5) (20)	29.83	7.67	93.90	209.65
PC(36:5) (79)	6.37	2.86	8.67	0.96
PC(36:5e)+PE(38:5e) (45)	8.40	2.93	15.53	4.67
PC(38:5) (18)	33.11	19.01	50.09	32.67
PC(38:5) (254)	1.03	0.65	1.47	0.02
PC(38:5) (64)	7.73	0.98	26.17	17.31
PC(38:5) (802)	16.72	8.41	24.75	8.16
PC(38:5e) (190)	1.45	0.71	2.92	0.21
PC(38:5e) (41)	10.54	5.33	17.45	5.44
PC(40:5) (72)	10.43	2.95	30.54	11.74
PC(40:5e) (231)	1.16	0.47	2.03	0.08
PC(40:5e) (742)	0.11	0.03	0.23	0.00
PC(34:6) (955)	0.24	0.10	0.49	0.00
PC(38:6) (414)	0.53	0.02	2.05	0.11
PC(40:6) (1033)	1.34	0.73	3.37	0.12
PC(40:6) (784)	0.17	0.08	0.74	0.01
PC(38:7e) (1135)	0.27	0.13	0.55	0.01
PC(40:7) (115)	4.48	2.69	6.28	0.37
PC(40:7) (199)	1.65	0.89	5.52	0.21
PC(36:8e) (107)	5.51	4.01	7.72	0.44
PC(36:8e) (1200)	66.71	34.50	109.48	125.44
PC(36:8e) (149)	2.42	0.39	7.84	1.08
PC(38:8e) (114)	4.71	2.95	7.06	0.54
PC(40:9) (124)	4.18	2.64	5.74	0.32
PC(40:10) (1193)	0.13	0.05	0.29	0.00
PC(40:10e) (1028)	3.49	1.25	5.28	0.51
LysoPC(14:0) (580)	0.95	0.35	2.08	0.10
LysoPC(16:0) (1111)	80.57	47.16	145.55	202.61
LysoPC(18:0) (58)	21.83	10.44	40.32	34.04
LysoPC(16:1) (355)	1.55	0.63	5.31	0.35
LysoPC(18:1) (66)	20.75	9.28	42.59	36.00
LysoPC(18:2) (56)	29.43	11.34	78.47	154.91
LysoPC(18:3e) (982)	0.27	0.10	0.59	0.01
LysoPC(20:3) (330)	1.99	0.85	5.14	0.47
LysoPC(18:4e) (625)	0.65	0.29	1.41	0.05
LysoPC(20:4) (192)	4.73	1.98	12.47	2.69
LysoPC(20:5) (534)	0.71	0.12	2.75	0.16
LysoPC(22:5) (657)	0.50	0.16	1.11	0.03
LysoPC(22:5) (919)	0.36	0.14	0.89	0.02
LysoPC(22:6) (394)	1.46	0.55	3.51	0.29

TABLE S 10: Description of Phosphatidylethanolamins and Lysophosphatidylethanolamins

Lipid (variable number)	Mean	Minimum	Maximum	Variance
PE(30:0e) (574)	1.05	0.35	2.52	0.19
PE(32:0) (646)	0.81	0.24	1.57	0.06
PE(32:0e) (1175)	32.91	20.42	51.37	47.85
PE(32:0e) (215)	8.74	4.86	15.86	4.32
PE(34:0) (101)	28.73	12.68	52.71	64.20
PE(34:0) (1119)	12.55	8.23	16.85	2.43
PE(34:0) (1246)	3.76	2.12	8.98	0.77
PE(34:0) (185)	8.70	4.74	19.88	4.46
PE(36:0) (513)	1.81	0.91	3.64	0.28
PE(32:1) (1074)	1.42	0.73	4.02	0.29
PE(32:1) (257)	5.42	2.84	11.95	2.80
PE(32:1) (713)	0.63	0.24	1.28	0.03
PE(34:1) (1050)	4.06	2.13	10.00	1.09
PE(34:1) (841)	1.53	0.41	10.95	1.36
PE(34:1e) (152)	11.56	6.49	27.12	8.91
PE(36:1) (145)	16.18	6.08	41.22	41.18
PE(36:1) (751)	0.81	0.24	4.69	0.26
PE(36:1) (823)	7.21	2.40	16.89	9.88
PE(36:1) (871)	1.69	0.44	10.68	1.36
PE(36:1) (92)	20.25	9.19	42.78	40.79
PE(36:1e) (422)	3.16	1.00	9.90	2.08
PE(38:1) (1030)	8.71	3.34	20.14	12.19
PE(34:2) (1334)	2.69	0.87	19.27	2.63
PE(34:2) (1335)	6.41	3.20	15.15	3.64
PE(34:2) (723)	0.60	0.22	3.73	0.08
PE(34:2e) (133)	14.57	6.82	26.04	12.44
PE(34:2e) (155)	13.07	6.47	26.78	14.09
PE(34:2e) (541)	0.64	0.25	1.41	0.03
PE(34:2e) (600)	1.09	0.56	1.65	0.05
PE(36:2) (90)	20.90	6.48	52.87	85.78
PE(36:2e) (408)	0.96	0.36	2.35	0.10
PE(38:2) (1286)	14.93	4.64	28.55	17.49
PE(38:2) (130)	16.32	5.17	37.13	38.89
PE(38:2) (222)	6.32	2.77	12.14	3.42
PE(32:3e) (652)	0.88	0.23	2.18	0.12
PE(36:3e) (279)	2.22	0.74	4.99	0.61
PE(38:3) (170)	8.40	3.09	21.49	9.28
PE(38:3) (189)	7.29	3.53	17.99	4.30
PE(38:3) (835)	7.92	3.52	22.43	5.41
PE(40:3) (299)	4.28	1.82	7.76	1.35
PE(40:3) (67)	33.74	16.18	71.80	88.82
PE(36:4) (1272)	36.22	20.47	51.05	27.72
PE(36:4) (510)	1.63	0.37	9.30	1.03
PE(36:4) (630)	0.90	0.43	3.97	0.17
PE(36:4e) (405)	1.38	0.55	3.37	0.26
PE(38:4) (1208)	2.71	0.72	13.91	2.10

PE(38:4) (1223)	27.86	13.79	71.54	62.92
PE(38:4) (243)	5.32	1.20	26.67	8.58
PE(38:4) (34)	56.42	27.68	148.60	243.88
PE(38:4) (606)	1.09	0.75	1.50	0.02
PE(38:4e) (597)	0.52	0.15	1.44	0.04
PE(40:4) (1100)	0.52	0.28	0.83	0.01
PE(40:4) (120)	14.81	6.52	25.91	13.00
PE(40:4) (325)	2.95	1.38	6.25	0.68
PE(40:4) (819)	15.99	6.95	30.62	17.42
PE(32:5e) (928)	0.71	0.32	1.95	0.05
PE(36:5) (464)	1.84	0.94	3.11	0.10
PE(38:5) (125)	12.54	4.46	40.33	36.33
PE(38:5) (202)	5.49	1.64	18.75	7.75
PE(38:5e) (91)	9.09	3.15	24.23	11.42
PE(40:5) (812)	24.49	13.26	42.50	32.35
PE(36:6e) (1178)	14.56	5.38	28.11	18.35
PE(36:6e) (252)	3.45	1.94	5.62	0.49
PE(36:6e) (692)	0.77	0.39	1.23	0.03
PE(38:6) (1117)	16.01	10.72	21.48	4.62
PE(38:6) (326)	3.77	2.66	5.71	0.27
PE(38:6e) (198)	4.20	1.56	9.82	2.50
PE(38:6e) (358)	1.47	0.35	5.12	0.63
PE(40:6) (1323)	3.59	1.21	9.80	2.01
PE(40:6) (217)	4.37	1.74	8.05	2.01
PE(40:6) (386)	2.43	0.60	5.89	0.56
PE(37:7e) (126)	17.09	10.00	26.08	9.64
PE(38:7) (1112)	84.95	45.04	126.95	169.38
PE(38:7) (117)	21.03	13.74	29.76	9.09
PE(38:7e) (289)	2.41	1.23	5.09	0.49
PE(38:7e) (375)	3.09	0.86	8.10	1.65
PE(38:7e) (485)	1.07	0.34	3.85	0.33
PE(40:7) (225)	7.28	2.46	16.52	4.57
PE(40:7) (88)	30.58	13.12	68.49	68.61
PE(40:7e) (265)	2.49	0.85	5.09	0.61
PE(40:7e) (518)	1.62	0.70	3.18	0.14
PE(37:8e) (357)	2.80	1.30	5.47	0.35
PE(40:8e) (123)	17.20	8.17	34.82	32.26
PE(40:8e) (151)	7.92	3.50	15.87	9.35
PE(40:8e) (560)	1.27	0.65	2.59	0.11
PE(40:8e) (609)	0.54	0.20	1.29	0.04
PE(39:9e) (127)	20.29	8.24	39.02	28.62
PE(39:9e) (1305)	381.13	263.95	521.22	2257.84
LysoPE(18:0) (545)	0.75	0.26	1.48	0.04
LysoPE(18:0) (554)	0.60	0.31	1.13	0.02
LysoPE(18:2) (368)	1.65	0.91	2.62	0.08

TABLE S 11: Description of Sphingomyelins

Lipid (variable number)	Mean	Minimum	Maximum	Variance
SM(d18:1/14:0) (118)	13.32	4.97	29.20	17.84
SM(d18:1/16:0) (13)	104.04	59.47	161.98	356.02
SM(d18:1/18:0) (61)	32.88	16.10	58.26	59.42
SM(d18:1/20:0) (52)	31.39	15.49	50.23	47.10
SM(d18:1/22:0) (186)	6.86	2.22	15.20	5.96
SM(d18:1/24:0) (825)	5.97	1.84	11.69	3.30
SM(d18:1/16:1) (434)	1.40	0.76	2.45	0.08
SM(d18:1/16:1) (68)	19.41	8.79	33.75	23.47
SM(d18:1/18:1) (1092)	0.61	0.20	1.54	0.05
SM(d18:1/18:2) (583)	0.80	0.31	2.29	0.13
SM(d18:1/24:2) (251)	2.19	0.25	6.10	1.04
SM(d18:1/25:2) (377)	1.55	0.69	3.23	0.18
SM(d18:1/26:2) (642)	0.49	0.25	1.09	0.02
SM(d18:1/18:3) (887)	1.29	0.47	3.57	0.24
SM(d18:1/25:3) (670)	0.43	0.21	0.69	0.01
SM(d18:1/27:3) (890)	2.03	0.84	4.69	0.30

Descriptive statistics of small polar metabolites and lipids excluded from *treelet transform* and principal component analysis due to high variances

TABLE S 12: Small polar metabolites that were excluded from the TT and PCA analysis (p=13), due to high variances (uppermost 10% of variances of log-transformed variables)

Small polar metabolite (variable number)	Variance of original variable	Variance of log-transformed variable
Erythrose* (1428)	367022415.0	22.4
Glucose* (1430)	57952268.2	19.5
Glucose* (1424)	343263382.0	19.3
Urea* (1395)	377075113.0	12.0
Glucose* (1418)	63425714.0	11.4
Urea* (1459)	65164789.0	11.3
Lactic acid* (471)	3852422.8	9.5
Hexanoic acid (599)	5780431.7	8.4
Gluconic acid, 2-oxo (553)	78534.5	7.2
Glycolic acid (365)	35512.4	6.4
2-Ethylhexanoic acid (527)	31459.4	5.7
Butanoic acid, 3-methyl-2-hydroxy (504)	120754.0	5.6
Glucose* (1394)	60992229.3	4.5

TABLE S 13: Lipids that were excluded from TT and PCA analysis (p=39) due to high variances (uppermost 10% of variances of original variables)

Lipid (variable number)	Variance of original variable
Cholesteryl ester	
ChoE(18:1) (50)	91.68
ChoE(20:4) (43)	147.79
ChoE(18:2) (17)	646.78
(Lyso)Phosphatidylcholines	
PC(32:1) (40)	66.88
PC(38:3) (15)	112.39
PC(36:8e) (1200)	124.88
LysoPC(18:2) (56)	154.23
LysoPC(16:0) (1111)	201.71
PC(36:5) (20)	208.73
PC(36:3) (8)	221.66
PC(38:4) (4)	392.10
PC(36:4) (1)	441.29
PC(34:2) (0)	450.08

Phosphatidylethanolamines	
PE(34:0) (101)	63.92
PE(40:7) (88)	68.31
PE(36:2) (90)	85.41
PE(40:3) (67)	88.43
PE(38:7) (1112)	168.62
PE(38:4) (34)	242.80
PE(39:9e) (1305)	2247.84
Sphingomyelin	
SM(d18:1/16:0) (13)	354.44
Triglycerides	
TG(56:7)* (1347)	65.70
TG(54:6) (1020)	70.10
TG(46:1) (184)	113.55
TG(50:0) (195)	115.95
TG(48:0) (147)	143.90
TG(48:2) (83)	158.81
TG(50:3) (30)	168.21
TG(54:5) (54)	187.45
TG(52:4) (21)	348.90
TG(54:2) (33)	371.63
TG(54:4) (22)	431.40
TG(52:1) (51)	539.26
TG(48:1) (48)	567.99
TG(54:3) (10)	631.66
TG(50:2) (14)	653.72
TG(52:3) (6)	1135.56
TG(50:1) (23)	1162.02
TG(52:2) (5)	1671.66

Cross-validation results from *treelet transform* on serum lipids

TABLE S 14: Results of cross-validations to find an optimal cut-level for tree generated with TT procedure on 353 serum lipids for 3 to 6 retained components

Nb of TCs	Cut-level 1	Cut-level 2	Cut-level 3	Cut-level 4
		(5 repetitions)		(10 repetitions)
3	327	327	326	325
4	327	327	325	325
5	324	323	323	323
6	323	325	322	324

Used STATA and SAS codes

STATA codes for *treelet transform*

STATA add on for *treelet transform* developed by Anders Gorst-Rasmussen [63]

Syntax for analysis on small polar metabolites

- tt ln_methionine2tms - ln_3phenylpropionicacidtms, cor cut(40) components(5)

  /generates the treelet dendrogram based on the correlation matrix of the stated metabolites, cuts the tree at level 40 and extracts 5 components/

- ttcv ln_methionine2tms - ln_3phenylpropionicacidtms, cor components(*)

  /executes cross-validation method to determine optimal cut-level; *number of components ranged from 3 to 6/

Syntax for analysis on lipids

- tt cerd181160l_707l - tg597l_1190l, cor cut(200) components(3)

  / generates the treelet dendrogram based on the correlation matrix of the stated lipids, cuts the tree at level 200 and extracts 3 components/

- ttcv cerd181160l_707l - tg597l_1190l, cor components(*)

  /executes cross-validation method to determine optimal cut-level, * number of components ranged from 6 to 6/

Additional syntax for both data sets

- ttdendro/ttscree

  /generates the distance matrix for the print of the dendrogram and prints the scree plot/

- ttstab

  /executes bootstrap analysis to assess stability of retained components/

- ttpredict/tt1score

  /results in prediction of component scores for further analysis/

SAS codes

SAS code for plotting dendrograms with distance matrix from STATA

```
*cluster metabolites according to their distances together;
proc cluster data=work.matrix_spmets (type=distance)
      method=single outtree=cluster_mets ;
      var methion_1 -- _3phenyl_121;
      id metabolite_name;
run;

*plot the corresponding tree;
goptions reset=all;
goptions htitle=0.5;

axis1 label=none value= (height=0.65 rotate=0 font=swiss);
axis2 value= (font = swiss height =0.6) label= (height=0.5)
label=none;
title 'Treelet dendrogram known small polar metabolites';
proc tree data=cluster_mets horizontal
    lines=(color=darkblue)
     vaxis=axis2
    haxis=axis1;
run;
```

SAS code for PCA on 121 serum metabolite variables

```
proc factor data=metab_low
method=prin /* principal component analysis (PCA) */
priors=one /*specifies prior communality estimates, users should
always specify prior=one when conducting a PCA */
scree     /* creates a plot that graphically displays the size of
eigenvalue associated with each component */
nfact=4   /*defines the number of factors to be retained*/
rotate=varimax  /* results in orthogonal (uncorrelated) components,
easier interpretable than correlated components */
round      /* causes all coefficients to be limited to two decimal
places */
flag=.35   /* mark every loading that exceeds .5 with an asterisk
(star symbol) -> makes it much easier to interpret
               a factor pattern; flag option should be used in
conjunction with the round option */
out=metab_low  /*defines the data set where the factor scores are
saved*/;
title 'PCA with known log transformed small polar metabolites with
low variances';
var metab_1_ln -- metab_2103_ln;/* variables which should be reduced
to a smaller number of principal components */
run;
```

Danksagung

Die vorliegende Arbeit wurde am Deutschen Institut für Ernährungsforschung in der Abteilung Epidemiologie erstellt. Auf diesem Weg möchte ich allen Personen danken, die mich in der Anfertigung dieser Arbeit unterstützt haben.

Besonderer Dank gilt meinem Betreuer Prof. Dr. Heiner Boeing für die Möglichkeit diese Arbeit im Rahmen der EPIC-Potsdam Studie anfertigen zu können, den großen Freiraum bei der Bearbeitung des interessanten Themas und für zahlreiche konstruktive Gespräche. Außerdem gilt mein Dank Prof. Dr. Reinhard Busse für die Betreuung meiner Promotion an der TU-Berlin.

Vielen Dank allen meinen Kollegen in der Epidemiologie für das stets angenehme Arbeitsklima und die viele Zeit in helfenden und statistisch weiterbringenden Gesprächen. Insbesondere meinen Doktoranden-Mitstreitern und allen Büro-Kollegen.

Ganz besonderer Dank gilt meiner Familie, die mich in meiner bisherigen Berufs- und Lebensgestaltung kontinuierlich unterstützt hat, mich jederzeit motivierte und damit grundlegend zum Gelingen dieser Arbeit beitrug. Spezieller Dank gilt meinem Freund Christian für das konstant offene Ohr und für viele hilfreiche Kommentare und Diskussionen.

Eidesstattliche Erklärung

Hiermit erkläre ich, die am Fachbereich Gesundheitswissenschaften der Technischen Universität Berlin eingereichte Dissertation mit dem Titel „*Treelet Transform* for untargeted metabolomics data: Investigation of serum metabolite and serum lipid components obtained with *Treelet Transform* and their association to anthropometry and intestinal microbiota in a sub-study of the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam“ selbstständig angefertigt und verfasst zu haben. Es wurden keine anderen als die angegebenen Quellen und Hilfsmittel benutzt.

Weiterhin versichere ich, die Arbeit an keiner anderen Hochschule oder Fachhochschule eingereicht zu haben.

Jana Förster **Berlin, 18.06.2014**