From the Institute of Genetic Epidemiology, Helmholtz Zentrum München
German Research Center for Environmental Health
Head: Prof. Dr. Strauch
and the Institute of Medical Informatics, Biometry and Epidemiology,
Faculty of Medicine, Ludwig-Maximilians-University Munich
Director: Prof. Dr. Ulrich Mansmann
Chair of Epidemiology: Prof. Dr. Dr. Wichmann (emeritus)

Phenotype Set Enrichment Analysis

Genome Wide Analysis of Multiple Phenotypes

Thesis
Submitted for a Doctoral Degree in Natural Sciences
at the Faculty of Medicine
Ludwig-Maximilians-University Munich

by
Janina S. Ried

from Vechta (birth place)

2012

Printed with approval of the medical faculty
of the Ludwig-Maximilians-University Munich

Supervisor/Examiner:

Prof. Dr. Dr. H.-Erich Wichmann

Co-Examiner:

Prof. Dr. Ulrich Mansmann

Co-Supervisor:

Dr. Christian Gieger

Dean:

Prof. Dr. med. Dr. h.c. Maximilian Reiser, FACR, FRCR

Date of oral examination:

11.06.2013

Für meine Eltern.

Summary

In the analysis of genetic causation of disease and genetic effects on quantitative traits genome wide association studies (GWAS) on single phenotypes are widely applied. Although it is known that the analysis of multiple phenotypes can help to identify new genetic loci and improve the understanding of the shared genetic determination of phenotypes, not many studies analyze multiple phenotypes in a combined way. In this thesis a new approach for the analysis of multiple phenotypes, named phenotype set enrichment analysis (PSEA), is presented and successfully applied to two different panels of phenotypes.

PSEA was developed to analyze genetic effects on sets of multiple phenotypes at a genome wide scale. These predefined sets are created using prior knowledge or criteria based on single phenotype associations. PSEA calculates per gene an *enrichment score* based on single phenotype associations. This score is tested for enrichment with a *permutation test*. The algorithm uses known methodological elements from gene set enrichment analysis on GWAS data that were adapted to the situation of phenotype sets. Apart from the basic algorithm two extensions were developed. One enables the internal identification of new phenotype sets that are tested for enrichment in addition to the predefined phenotype sets. With this approach, PSEA can use prior knowledge and generate new hypotheses. A second extension allows the usage of GWAS summary statistics instead of individual level genotypes and phenotypes. The algorithm was implemented in C with MPI (message passing interface) parallelization and executed on a high performance computer. It was applied to a panel of iron related phenotypes and blood count traits using data of the population based KORA studies (cooperative health research in the region

of Augsburg). PSEA identified additional genetic loci that had an effect on blood and iron traits and were not genome wide significant in single phenotype GWAS on the same data. These findings were confirmed by results of published large meta-analyses. At second, PSEA was applied to two panels of metabolite measurements including more than hundred metabolites each. Predefined phenotype sets were determined with a Gaussian graphical modeling (GGM) approach. It was shown that PSEA can deal with large panels of more than 150 phenotypes. Various findings from single phenotype GWAS on the same data were confirmed. The internal identification of new phenotype sets unraveled several unknown shared genetic effects and provided valuable insight into the network of metabolites.

Both applications showed that PSEA is able to identify additional loci that were not found in single phenotype GWAS and improve the knowledge on phenotype relation and underlying genetic basis. Moreover, the algorithm is flexible, requires minimal assumption and can easily be applied in various situations.

Zusammenfassung

In der Analyse der genetischen Ursachen von Krankheiten oder Beeinflussung quantitativer Parameter werden häufig genomweite Assoziationsstudien (GWAS) einzelner Phänotypen verwendet. Obwohl die Analyse von multiplen Phänotypen zur Identifikation neuer genetischer Loci führen kann und zudem das Verständnis gemeinsamer Effekte auf verschiedene Phänotypen erweitert, werden in den meisten Studien Phänotypen separat ausgewertet. In dieser Arbeit wird eine neue Methode, "Phenotype Set Enrichment Analysis" (PSEA), vorgestellt und die vielversprechenden Ergebnisse ihrer Anwendung auf zwei verschiedene Phänotyp-Datensätze präsentiert.

PSEA analysiert genomweit den Zusammenhang genetischer Loci und multipler Phänotypen, die zu Phänotyp-Gruppen (phenotype sets) zusammengefasst werden. Die vordefinierten Phänotyp-Gruppen werden unter Verwendung von bestehendem Wissen oder Ergebnissen aus einzelnen Phänotyp Analysen generiert. PSEA berechnet aus den Assoziationen der einzelnen Phänotypen mit einer Gen Variante den sogenannten *Enrichment Score*. Die Assoziationen einzelner Phänotypen mit der Gen Variante werden mit diesem *Score* erfasst. Mit einem *Permutationstest* wird getestet, ob der Score höher ist als zufällig erwartet. Für den Algorithmus wurden Elemente bekannter Methoden der Gen-Gruppen Analyse (gene set enrichment analysis) in angepasster Form verwendet. Zusätzlich zum grundlegenden Ansatz wurden zwei Erweiterungen entwickelt. Zum einen eine Strategie zur Identifikation neuer Phänotyp-Gruppen mit PSEA. Die neu identifizierten Phänotyp-Gruppen werden auf dieselbe Art analysiert wie die auf Grundlage von vorherigem Wissen definierten Mengen. Somit besteht mit dieser Erweiterung die Möglichkeit mit PSEA nicht nur Hypothesen bestehenden Wissens zu testen sondern auch

neue Hypothesen zu generieren. Die zweite Erweiterung ermöglicht die Verwendung von GWAS Ergebnissen anstelle von Genotyp- und Phänotyp-Messungen pro Person. PSEA wurde in C implementiert, mit MPI (message passing interface) parallelisiert und auf einem Hochleistungs-Rechencluster ausgeführt. Die erste Anwendung von PSEA untersuchte quantitative Phänotypen, die mit dem Eisenmetabolismus oder dem Blutbild zusammenhängen. Dabei wurden zusätzliche genetische Loci identifiziert, die einen Effekt auf mehrere Blut- und/oder Eisenphänotypen zeigen, aber in GWAS der einzelnen Phänotypen auf derselben Datengrundlage keinen genomweit signifikanten Effekt zeigten. Publizierte GWAS und Metaanalysen großer Studien identifizieren ebenfalls diese Loci und bestätigen somit die Ergebnisse, die mit PSEA auf Grundlage einer geringeren Anzahl Individuen gefunden wurden. Als zweite Anwendung von PSEA wurden zwei umfassende Datensätze von jeweils mehr als hundert Metaboliten analysiert. Die Phänotyp-Gruppen wurden mit Gaußschen graphischen Modellen (GGM) bestimmt. Ergebnisse bestätigten Assoziationen zwischen Gen Varianten und verschiedenen Metaboliten, die aus den einzelnen GWAS der Phänotypen bekannt waren. Die interne Identifikation neuer Phänotyp-Gruppen ermöglichte die Identifikation bisher unbekannter genetischer Zusammenhänge und neue Erkenntnisse über das Zusammenspiel der Metaboliten.

Beide Anwendungen haben gezeigt, dass PSEA genetische Loci identifizieren kann, die bei der separaten Analyse der Phänotypen nicht gefunden wurden, und neue Erkenntnisse über die gegenseitige Verflechtung der Phänotypen mit der zu Grunde liegende genetischen Kondition liefert. Zudem ist der Ansatz flexibel, kommt mit minimalen Annahmen aus und kann einfach auf unterschiedliche Situationen angewendet werden.

Contents

List of Figures

List of Tables

List of Abbrevations

cis-eQTL	cis-expression quantitative trait locus
ES	enrichment score
FDR	false discovery rate
FWER	family wise error rate
GGM	Gaussian graphical modeling
GWAS	genome wide association study
kb	kilo base pairs
KORA	cooperative health research in the region of Augsburg
LD	linkage disequilibrium
NES	normalized enrichment score
PSEA	phenotype set enrichment analysis
SNP	single nucleotide polymorphism

1. Introduction

Direct-to-consumer genetic test are controversially discussed in the scientific community (Sivakumaran *et al.*, 2011; Wade and Wilfond, 2006; Borry *et al.*, 2010). The services offered by various companies provide genome wide or disease specific genetic testing. One critical aspect of such tests is that genes can influence different traits and additional effects might be found in the future. Therefore, testing a genetic risk for a specific disease or trait might lead to unintended information about another additional disease (Sivakumaran *et al.*, 2011; Wade and Wilfond, 2006). This could lead to ethical problems, especially if the second disease is untreatable or life threatening. This is only one example for the importance of further insight in genetic effects on multiple traits.
It is biological plausible that genes can have effects on different phenotypes. Gene products participate in various biological processes and one molecular function of a gene can have multiple consequences (He and Zhang, 2006). A recent study finds that approximately 17 % of all genes that are known to be associated with diseases or disease traits have an effect on multiple (independent) phenotypes (Sivakumaran *et al.*, 2011). In contrast to that joint analyses of multiple phenotypes are not often performed at the moment. Mostly, genome wide association studies (GWAS) of single traits are applied for analysis of genetic effects on quantitative phenotypes and diseases. Some studies present the results for several related phenotypes side by side in a comparing manner (Soranzo *et al.*, 2009), but most published GWAS are limited to single quantitative phenotypes or one disease related trait (e. g. Kilpeläinen *et al.*, 2011; Oexle *et al.*, 2011).

This thesis presents a new method for the analysis of multiple phenotypes that analyzes in parallel various sets of phenotypes for gene based association on a genome wide scale.

The understanding of the relation of genes and different traits or diseases is one aim of analyzing multiple phenotypes. Investigation of the underlying structure of different phenotypes and genes can give important insights into biological mechanisms and connections. For example, the iron metabolism and the hematopoiesis are known to be connected (Orkin and Zon, 2008). A joint analysis of iron and blood related traits can improve the understanding of shared pathways. For large panels of intermediate phenotypes like metabolomics a combined analysis can help to understand the hidden connections and give insight into the function of metabolites and genes.
The second aim of multiple phenotype analysis is the identification of new loci. If the analysis is focused on one phenotype and more phenotypes are available, only a part of the information is used. The shared information of correlated phenotypes can be advantageous for these analyses. For example, if a gene is associated with several phenotypes, but the effect is too small to be detected in single phenotype GWAS, the association could possibly be identified if the phenotypes are analyzed together. In other words, using more phenotypes is equivalent with using a larger fraction of the available information.

The methods for multiple phenotype analysis are under development at the moment, proposing various approaches (Shriner, 2012). Some authors developed multivariate algorithms (Ferreira and Purcell, 2009; Liu *et al.*, 2009) that are of limited use if many phenotypes are under consideration. Moreover, they cannot be applied easily to situations where phenotype and genotype measurements for each person are not available. Other approaches use the results of GWAS on single phenotypes (Gupta *et al.*, 2011; Huang *et al.*, 2011; Yang *et al.*, 2010). The focus of these algorithms is hypothesis generation, mining the data for sets of phenotypes associated with genetic loci. Another strategy is to perform a phenotype screening for a previously defined genetic locus, as done in the phenome wide association scans (PheWAS) (Denny *et al.*, 2010).
A flexible approach for the genome wide analysis of multiple phenotypes that can deal with large number of phenotypes, easily integrate prior knowledge, generate new knowledge and manage minimal assumptions is lacking.

This work describes a new strategy for the analysis of multiple phenotypes that satisfies these requirements. The method tested phenotypes combined in phenotype sets for enrichment at genes. Current methods for multiple phenotypes address either single nucleotide polymorphism (SNP) level (Yang *et al.*, 2010) or gene level (Huang *et al.*, 2011) effects on multiple phenotypes. As genome wide scan calculations based on SNP level imply a large multiple-testing burden and require a consideration of the underlying LD structure, this work was focused on the gene based approach. The gathering of phenotypes in phenotype sets had the advantage that prior knowledge could easily be integrated and not many assumptions had to be made, except an approximately normal distribution of the phenotypes. For the enrichment test methodological elements known from gene set enrichment analysis approaches (Ackermann and Strimmer, 2009; Wang *et al.*, 2010) were adapted to the specific situation of investigation of phenotype sets.
In a nutshell, for each phenotype set and gene variant a score was calculated as a combination of the statistics of the corresponding univariate analyses of all phenotypes in the phenotype set. Then the enrichment of this score was evaluated by a permutation test. An extension allowed the identification of new phenotype sets. Another extension enabled the usage of summary statistics of single phenotype GWAS instead of individual phenotype and genotype measurements per person. This algorithm, named phenotype set enrichment analysis (PSEA), is the first adaption of a set enrichment approach to the analysis of multiple phenotypes.

The method was applied to two different studies of the cooperative health research in the region of Augsburg (KORA): KORA F3 and KORA F4. The KORA cohorts are several cohorts representative of the general population in Augsburg, Southern Germany and two surrounding counties that were initiated as part of the world health organization (WHO) monitoring cardiovascular disease (MONICA) study (Wichmann *et al.*, 2005). From the collected data genome wide genotyping, phenotypes related to iron metabolism, blood cell counts and metabolite measurements were used. Replication analysis was performed in TwinsUK data, which is an adult twin-registry (Andrew *et al.*, 2001).

In this thesis the development of the PSEA algorithm and its application to two different panels of phenotypes is described. Chapter two outlines the general method of PSEA and two extensions. In chapter three the application to iron related phenotypes and blood count traits is presented. The fourth and fifth chapter report the results of an application of PSEA to two large panels of metabolomics measurements (Biocrates and Metabolon) using two different kinds of metabolite set definition. The findings are summarized in chapter six and discussed in terms of possibilities, limitations and further applications.

The general method of PSEA is pictured in chapter two. Therefore, the algorithm is described in four main steps: calculation of gene based test statistics, calculation of the enrichment score, permutation test and estimation. Subsequently, two extensions of PSEA are presented. Those enabled the identification of new phenotype sets and the usage of GWAS summary statistics instead of individual phenotype and genotype measurements. The chapter is closed by a discussion of some important methodological elements of PSEA.

The third chapter deals with the application of PSEA to a panel of iron related phenotypes and blood count traits. For this selection of phenotypes it could be exploited that many genes were known from previous publications to be associated with at least one of these phenotypes as this knowledge was used for the evaluation of the results of PSEA. The application of PSEA identified several significant enrichments in KORA F4 that could be replicated in KORA F3. Comparison with results from large meta-analyses confirmed the identified phenotype set enrichments. PSEA was able to find additional loci that could not be detected in single phenotype GWAS on the same data. The extension of PSEA on GWAS summary statistics was applied to the same phenotypes using GWAS summary statistics instead of individual genotype and phenotype measurements. The results were similar to the original approach of PSEA on genotype and phenotype measurements per person, but the extension resulted higher p-values. Reasons for this observation are evaluated. For all results biological interpretation with respect to published studies is given. Moreover, this outcome is discussed more generally as a *Proof-of-Principle* of the PSEA strategy.

The application of the general algorithm of PSEA to two large sets of metabolite panels is described in chapter four. The metabolites were measured with two techniques provided by Biocrates (BIOCRATES Life Sciences AG) and Metabolon for individuals of KORA F4, used for discovery, and TwinsUK, used for replication. Metabolite sets were determined with a Gaussian graphical modeling (GGM) approach that exploits the partial correlation structure of the metabolites. Several metabolite sets were found to be significantly enriched and replicated. The results corresponded to biological plausible relations of gene function and its (by-)products and substrates. Known associations from single phenotype GWAS were confirmed. The chapter is closed by a discussion of methodological aspects of the application of PSEA to metabolite panels and metabolite set definition.

The fifth chapter describes the application of the extension for identification of new phenotype sets to the same metabolite data as in chapter four. Several new genetic loci were found for an enrichment with new identified phenotype sets. Additional information could be gained for known loci due to new identified phenotype sets. In the end of this chapter the extension of PSEA and its findings are discussed with respect to the situation of metabolomics data.

The last chapter summarizes the findings of PSEA and discusses its strengths and possible limitations. Ways for integrating PSEA in an analysis pipeline are mentioned. Ideas for modifications and extensions of PSEA are given as a perspective on further development possibilities.

2. Phenotype Set Enrichment Analysis (PSEA)

The aim of PSEA is to identify associations between genes and a set of phenotypes. The general method is a phenotype set based enrichment approach. This chapter describes the development of PSEA using adapted elements of established gene set enrichment approaches. First, the general algorithm for testing the enrichment of a phenotype set at a gene is described. Afterwards, two extensions are presented, namely the identification of new phenotype sets that are likely to be enriched for a gene and the use of GWAS summary statistics instead of phenotype and genotype measurements per person. The chapter is closed by the critical reflection of some basic concepts of PSEA.
The algorithm of PSEA was recently published: Ried *et al.* (2012).

2.1. PSEA – General Algorithm

For the algorithm methods already used for gene set enrichment analysis and its application to GWAS data (Efron and Tibshirani, 2007; Guo *et al.*, 2009; Segrè *et al.*, 2010; Subramanian *et al.*, 2005; Wang *et al.*, 2007) were modified. Parallels of gene set enrichment analysis and phenotype set enrichment analysis facilitated usage of the same methodological approaches. Anyway, specific characteristics of the analysis of phenotype sets had to be regarded. In the following description of the algorithm it is stated for each step which elements were borrowed from gene set analysis approaches and how they were modified. A schematic overview of the main steps of PSEA is given in Figure 2.1.

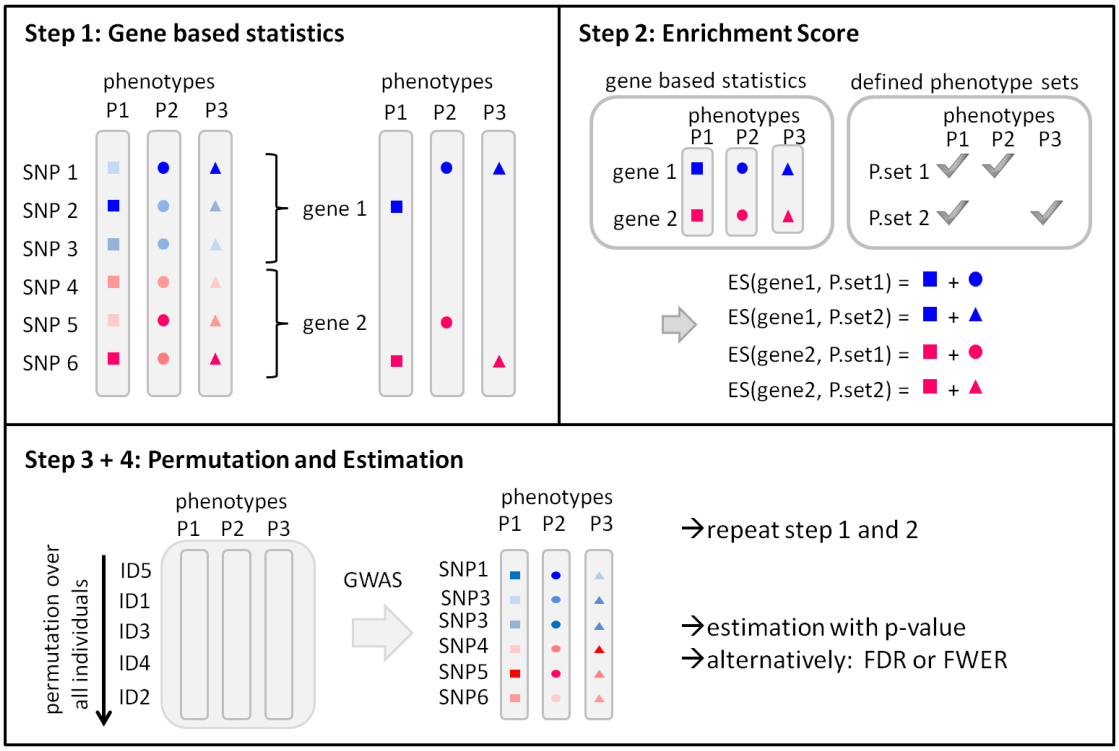


Figure 2.1.: **Schematic overview of the PSEA algorithm.** P1, P2, P3 are phenotypes, P.set1 and P.set2 denote phenotype sets. The strength of association of a SNP or gene with the phenotypes is indicated by different shadings of red and blue, a darker colour representing a stronger association.

In the following p_m with $m \in \{1, \ldots, N_{pheno}\}$ is one phenotype of N_{pheno} continuous phenotypes, which are approximately normally distributed, and PS is a phenotype set that is a selection of N_{PS} phenotypes from p_1,...,p_{Npheno}.

Step 1: Gene Based Statistics

To test genes for enrichment of a phenotype set, a gene based statistic was required for each phenotype. The positive test statistics of an association test for all SNPs mapped to a gene were calculated (e.g. χ^2-test statistic for testing the effect estimate in a linear regression). These statistics were combined to a gene-wise test statistic per phenotype. Analogous to gene set enrichment methods for GWAS data (Guo *et al.*, 2009; Wang

et al., 2007), the maximal test statistic of all SNPs mapped to a gene was taken as the gene based statistic $t(gene, p_m)$. All SNPs in the transcribed region and a surrounding area of a 110 kilo base pairs (kb) upstream and 40 kb downstream were mapped to the gene. This definition for assigning SNPs to genes was chosen as 99 % of the expected cis-expression quantitative trait loci (cis-eQTLs) are located within this interval (Segrè *et al.*, 2010; Veyrieras *et al.*, 2008).

Step 2: Enrichment Score

The next step was the determination of the enrichment score (ES) for each combination of phenotype set and gene. The ES was calculated as sum of the gene based statistics of all phenotypes in the phenotype set:

$$ES(PS, gene) = \sum_{p_m \in PS} t(gene, p_m) \ . \tag{2.1}$$

This is a modification for positive test statistics of the widely used *maxmean* statistic (Efron and Tibshirani, 2007), which was presented as *sumstat* score for gene set enrichment (Tintle *et al.*, 2009).

Step 3: Permutation

For PSEA a so called *self-contained* test was applied. The self-contained approach tests whether a phenotype set is enriched compared to sets of random phenotypes (Ackermann and Strimmer, 2009; Wang *et al.*, 2010). It ensures that the result of PSEA for one phenotype set is independent of other phenotype sets under consideration (more details in the section 2.4).

If individual level phenotypes and genotypes are available, random phenotypes can be generated by permutation of the values over all individuals. This was also proposed for different gene set enrichment approaches (Wang *et al.*, 2007). A special requirement for PSEA is that one has to preserve the correlation structure of the phenotypes. Otherwise, sets of phenotypes and sets of permutated phenotypes would not be comparable. Therefore, all phenotypes were permuted in the same way. An algorithmic formulation

of the permutation scheme is given in the Appendix A.1. Based on each permutation of the phenotypes the SNP-wise test statistics could be calculated. Analogous to the calculations for the original phenotype values, gene based statistics per gene (compare step 1) and the permutation ES (compare step 2) were derived

$$ES^{(j)}\left(PS, gene\right) = \sum_{p_m \in PS} t\left(gene, p_m^{(j)}\right) . \tag{2.2}$$

Thereby, the index (j) indicates a permutation with $1 \leq j \leq N_{perm}$.

For estimations, where it is important that the ES were comparable for different genes and phenotypes sets, a normalized enrichment score (NES) can be used. This was calculated by normalizing the ES with the mean and standard deviation of the ES for all permutations (see Appendix A.2 for details).

Step 4: Statistical Significance and Multiple Testing Corrections

Statistical testing was performed using empirical distributions generated by permutation of phenotype values. Various methods have been applied for gene set enrichment analysis. For example a permutation p-value can be used for estimation in PSEA.

Permutation p-value: The permutation p-value of the test if a phenotype set (PS*) is enriched for a gene (gene*) was taken as the fraction of permutations for which the permutation based ES is larger than the original ES

$$P\left(PS^*, gene^*\right) = \frac{\#_j \left[ES\left(PS^*, gene^*\right) \leq ES^{(j)}\left(PS^*, gene^*\right)\right]}{N_{perm}} . \tag{2.3}$$

Thereby, the index j indicates the permutation ($1 \leq j \leq N_{perm}$) and $\#_j$ means count over all permutations. The significance level needed to be corrected for the number of phenotype sets times the number of genes. As the number of permutations was limited, the p-value was not continuous but raised in steps of $1/N_{perm}$.

Alternatively, the false discovery rate (FDR) and family wise error rate (FWER) could be used (see Appendix A.2 for details).

Implementation

The algorithm of PSEA was implemented in the programming language C with MPI (message passing interface) parallelization. It was designed as a flexible command line based software. Phenotype data, information about phenotype sets and gene mapping were processed from human readable files in plain text format. Genotypes in Impute output format (Marchini *et al.*, 2007) could be used as well as zipped or binary Impute output format. The parallelization enabled calculation of enrichment for multiple genes in parallel. For example the software was executed within the DEISA (Distributed European Infrastructure for Supercomputing Applications) project on a supercomputer with 400 processes in parallel (see section 3.3).

2.2. Extension 1: Identification of New Phenotype Sets

The basic algorithm of PSEA as described above tested only predefined sets that were specified by prior knowledge and used as input for PSEA. But if only predefined phenotype sets are used, one could miss an important combination of phenotypes. This is especially true as different phenotype sets may be enriched for different genes. Testing all possible combinations of sets would raise the number of tests and is therefore often not feasible. The identification of new phenotype sets parallel to the testing of predefined sets enabled to analyze promising associations.

For each gene a threshold criterion based on single phenotype association results was applied. All phenotypes that had gene based univariate test statistics higher than a predefined threshold were regarded as newly identified phenotype set for this gene. A p-value threshold of $5 \cdot 10^{-4}$ was applied. This threshold was used only for identification of a phenotype set. Testing for enrichment of the new identified set was done in the same way as for the predefined sets. Therefore, the usage of a threshold that was less stringent than the univariate significance level was permitted. This aimed to balance between including phenotypes on which the gene has no effect and missing phenotypes on which the gene has an effect.

2.3. Extension 2: PSEA on GWAS Summary Statistics

A modification of PSEA was developed that can use test statistics per SNP and phenotype from GWAS results if phenotype and genotype measurements for each person are not available. The permutation strategy was the major element that had to be adapted. Instead of permuting phenotype levels over all individuals, the test statistics were permuted over all SNPs. Guo *et al.* (2009) proposed a similar approach for gene set enrichment analysis. To preserve the correlation structure of the phenotypes, the vector of the SNP-wise test statistics was permuted in the same way for all phenotypes. Gene based test statistics and permutation ES were calculated as described above. Testing was performed with p-value estimation (as presented in equation (2.3)).

There are two impacts that have to be considered for this permutation scheme: (A) the distribution of test statistics might be influenced by associations of the phenotypes with other genes and (B) the permutation of SNP-wise test statistics destroys the linkage disequilibrium (LD) structure. In other words, (A) means that the vector of SNP-wise test statistics might include more high test statistics than expected under the distribution of the null hypothesis of no SNP phenotype association. In permutations these high test statistics could lead to higher permutation ES and therefore reduce the number of identified enrichments. Of course, this is dependent on the phenotypes in the considered phenotype set, especially on their strength of genetic associations and number of associated SNPs. The effect of the destroyed LD structure in the permutations could also lead to higher permutation ES. That is because of the gene test statistics, which were calculated as maximum of all SNP test statistics. By chance, the maximal test statistic of independent SNPs will be higher than the maximal test statistic of dependent SNPs. The effect of these two impacts was studied exemplarily for the data of the first application (see section 3.3).

2.4. Discussion of Methodological Design

Level and Power of PSEA are dicussed in the Appendix A.3

Effect of SNP-Gene Mapping

As PSEA is a gene based approach, a SNP-gene mapping is needed. Such a mapping defines which SNPs are analyzed as one locus in PSEA. Above, the strategy was proposed to map all SNPs to a gene that are in the transcribed or a flanking region of 110 kb upstream and 40 kb downstream. This SNP-gene mapping was designed to cover not only the transcribed region but also most cis-eQTLs (Segrè *et al.*, 2010). Therefore, it is reasonable in terms of creating a good representation of each gene. Due to the overlapping definition of genes, the loci are not independent. To regard this overlapping structure of genes in the evaluation of PSEA results, one can consider groups of genes that are overlapping in at least one SNP (gene groups). For non genome wide applications of PSEA the SNP-gene mapping could be reduced to independent genes of interest. Moreover, PSEA can be applied with any other SNP-gene mapping if there are good reasons for a modification.

Apart from different SNP-gene mapping methods, one could also extend PSEA on SNP level. Technically, this approach would be possible analogous to the gene based analysis by using single SNP statistics. There are two main situations where the SNP based PSEA could gain more information than the gene based PSEA: First, PSEA on SNP level enables to find multiple independent loci that have an effect on a phenotype set in one gene. Second, SNPs with an effect on a phenotype set could be identified, even if they are not mapped to a gene. The main drawback of the SNP level calculation is that it would strongly increase the number of tests. Moreover, many SNPs are in high LD and therefore dependent of each other. To get reliable results, it would be reasonable to reduce the analysis to independent SNPs. But the additional computational effort (more tests, LD analysis) might be higher than the possible findings. As an extension, it is thinkable to use SNP level PSEA for analyzing genes for which gene level PSEA identified significant enrichments.

Effect of Permutation

The design of the permutation strategy has an important effect on the results of PSEA. It is obvious that the number of permutations has a direct impact on the PSEA results. Especially, the significance level for the p-value criterion is dependent on the number of permutations as it determines the gradation and therefore the lowest possible p-value. For application of Bonferroni-correction for PSEA the significance level has to be adjusted for the number of genes and the number of phenotype sets under consideration. That means for a genome wide run with approximately 20,000 genes, only one predefined phenotype set and 1,000 permutations the Bonferroni-corrected significance level $\frac{0.05}{20{,}000} = 2.5 \cdot 10^{-6}$ is below the lowest possible p-value. It is not always feasible to calculate a number of permutations that allows Bonferroni-correction, due to limitations in computational ressources. The usage of a lower number of permutations can lead to false positive findings.

A replication step helps to reduce the number of false positive findings that may result from the higher significance level. Another possibility is to increase the number of permutations stepwise. Such a strategy starts with a low number of permutations (e.g. 100) for all genes and increases the number of permutations only for those genes for which at least one phenotype set showed a promising low p-value for the first permutations. Several steps can be performed unless a sufficient number of permutations is achieved or the limitations of computational resources are reached. The results gained from the final step, the highest number of permutations, can be analyzed in a replication step.

One important aspect of the permutation strategy is that the correlation of phenotypes is conserved. In the basic algorithm of PSEA on individual level genotypes and phenotypes the phenotypes were permuted over all individuals. For application of PSEA to GWAS results the test statistics of each phenotype were permuted over all SNPs. In both cases all phenotypes were permuted in the same way so that the correlation structure of the phenotypes was conserved. The conservation of correlation of phenotypes in the permutation is important as the correlation of the phenotypes in the phenotype set has an effect on the ES and permutation ES. The permutation test compares the ES

of phenotype sets based on original and permuted phenotypes. Different correlation structures in the original data and its permutation would change the results. In other words, the permutation strategy ensures that the result for a phenotype set is not affected by modified correlation structure of the phenotypes in the set.

Self-Contained Test Strategy

It was mentioned above that PSEA follows a self-contained test strategy. That means it tests a phenotype set for enrichment at a gene by comparison with a phenotype set of random phenotypes. Each test for a phenotype set is independent from other phenotypes (e.g. from other phenotype sets). The applied formula for the ES and permutation scheme were designed to support this independence.

In contrast to that, in gene set enrichment often a *comparative* or *nested* approach is applied (Ackermann and Strimmer, 2009; Wang *et al.*, 2010). These strategies compare the enrichment of a set with the enrichment of a random set of all elements under consideration. Transferred to the phenotype set enrichment, a comparative approach would compare a set of phenotypes with random sets of all phenotypes under consideration and not with a set of random phenotypes. To gain reliable results, the comparative and nested strategy needs a large panel of independent set elements. For gene set enrichment often many independent genes are available but for phenotype set enrichment the number of phenotypes is limited and phenotypes are often correlated. To ensure reliable results, that are replicable in other studies with different phenotype availability, it is important that the enrichment analysis of a phenotype set is independent from other phenotypes in the analysis. The application of the self-contained strategy warrants this and is therefore the appropriate strategy for PSEA.

Other Approaches

PSEA belongs to the methods that combine statistics of univariate analysis. In comparison with multivariate analysis such approaches require in general fewer assumptions about the phenotypes (Yang *et al.*, 2010). Therefore, they can be transferred more easily to different situations such as the use of either categorial or continuous phenotypes.

PSEA could easily be modified for the analyses of binary traits or combinations of binary and quantitative traits. Besides PSEA, at least two other algorithms of that kind have been published recently (Huang *et al.*, 2011; Yang *et al.*, 2010). Yang *et al.* (2010) proposed a variation of O'Briens method (O'Brien, 1984) to combine univariate GWAS results, which was realized on SNP basis in contrast to our gene based approach. The program PRIMe (Huang *et al.*, 2011) identifies pleiotropic regions by scanning GWAS results of multiple phenotypes for low p-values, whereupon the LD structure is taken into account. PRIMe can identify different phenotype sets for different genes but it is not designed for testing a panel of predefined phenotype sets. Therefore, prior knowledge cannot easily be integrated in the analysis.

PSEA is able to use prior knowledge and generate new knowledge. Moreover, it is easy to apply and can deal with many phenotypes. These advantages are demonstrated in the applications presented in the following chapters.

3. Application to Iron and Blood Phenotypes

In this chapter the application and results of PSEA to a panel of blood count traits and iron related phenotypes are described. These phenotypes were a useful example to demonstrate the abilities of PSEA, as several of these phenotypes are known to be connected via the genesis of red blood cells. The fact that many genes were previously known to be associated with at least one of these phenotypes (Figure 3.1) was exploited to evaluate the results of PSEA. First the data used for the analysis is described. Then the results of the application of PSEA are presented. After that the results will be discussed not only with regard to their biological interpretation but also as a *Proof-of Principle* of PSEA.
This application was recently published together with the general algorithm of PSEA: Ried *et al.* (2012).

3.1. Iron and Blood Traits

PSEA was applied to genotype and phenotype data of two cohorts in the KORA study. The KORA study consists of several cohorts representative of the general population in Augsburg, Southern Germany and two surrounding counties that were initiated as part of the WHO MONICA study (Wichmann *et al.*, 2005). Ten years age-sex strata have been sampled from the 25 to 74 year old population with a stratum size of 640 subjects. In the KORA S3 study 4,856 subjects (response rate 75 %), and in KORA S4 in total

4,261 subjects have been examined (response rate 67 %). 3,006 individuals participated in a follow-up examination of S3 in 2004/05 which is called KORA F3. Follow-up for the S4 survey was performed in 3,080 individuals in 2006/2008 (KORA F4). All study participants underwent a standardized face-to-face interview by certified medical staff and a standardized medical examination including blood draw and anthropometric measurements. For discovery a cohort of 1,814 randomly sampled, unrelated individuals of the population based cohort KORA F4 was used. Replication was performed in a random sample of 1,644 unrelated individuals of the independent cohort KORA F3 (Wichmann *et al.*, 2005).

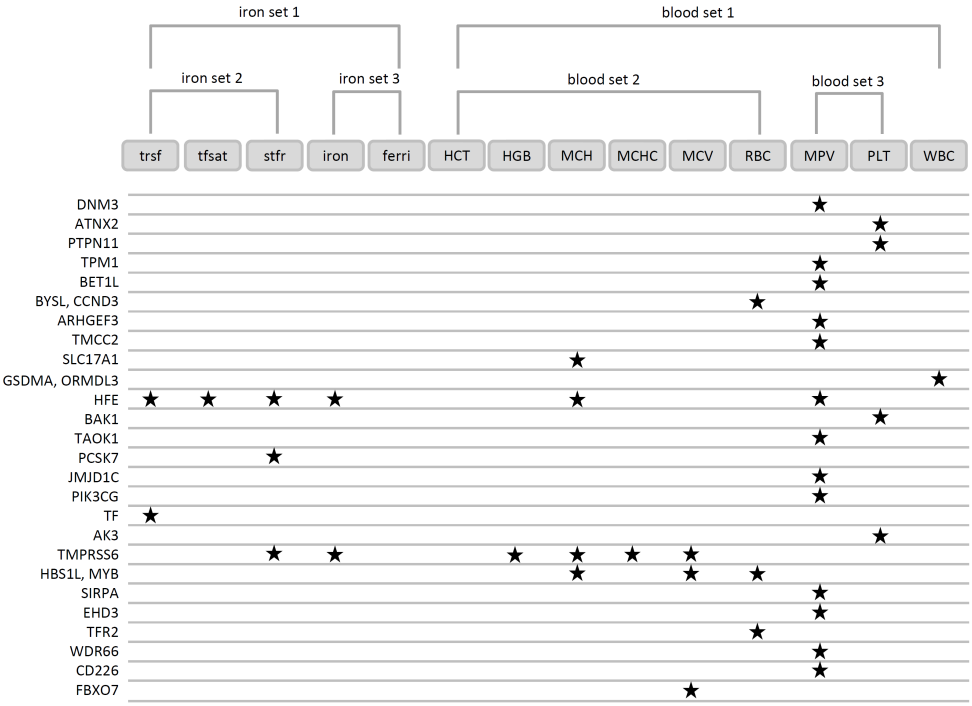


Figure 3.1.: **Iron and blood trait phenotype sets with published genes.** This figure summarizes the predefined phenotype sets analyzed with PSEA. All genes that have been reported to be significantly associated with at least one phenotype in previous GWAS or meta-analysis including multiple cohorts are indicated. (trsf: transferrin, tfsat: transferrin saturation, stfr: soluble transferrin receptor, ferri: ferritin, other abbreviations as given in the text)

PSEA, using phenotype and genotype measurements per person, required four types of input data: definition of phenotype sets, phenotype values for at least all elements of the phenotype sets, genotypes, and a SNP-gene mapping. For the application of PSEA to GWAS results one would not need phenotype values and genotype data but summary results from the GWAS for all phenotypes under consideration.

Phenotypes: The method was applied to a set of fourteen phenotypes (Figure 3.1): Five traits related to the iron metabolism (iron, ferritin, transferrin, transferrin saturation, soluble transferrin receptor) and nine traits related to blood cells including six red blood cell traits (haematocrit (HCT), haemoglobin (HGB), mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC), mean corpuscular volume (MCV), red blood cell count (RBC)), one white blood cell trait (white blood cell count (WBC)) and two platelet traits (mean platelet volume (MPV), platelet count (PLT)) (see appendix for details on measurement methods (Table B.1) and population statistics (Table B.2)). Outliers were excluded if they differed more than three standard deviations from the mean value. Residuals of a linear regression on each log-transformed phenotype concerning age and sex were calculated and taken as phenotypic input.

Phenotype Sets: Six phenotype sets were tested for enrichment; three included various combinations of iron traits and three included combinations of blood traits (Figure 3.1). These sets were selected based on obvious relation (e.g. white blood cells vs. red blood cells) and prior knowledge (iron intake and accumulation: iron and ferritin, iron transport: transferrin, transferrin saturation and soluble transferrin receptor (Oexle *et al.*, 2011; Hentze *et al.*, 2010)).

Genotypes: Genotyping for KORA F3 was performed with Affymetrix 500K genome wide SNP-array. The KORA F4 individuals were genotyped on Affymetrix 6.0 SNP-array. After SNP-wise and person-wise filtering genotypes were imputed with Impute v 1.0.0 (KORA F3)/Impute v 0.4.2 (KORA F4) (reference HapMap phase 2, release 22) (Marchini and Howie, 2010). The analysis was restricted to SNPs that had a minor allele frequency higher than 5 %, call rate higher than 95 % and imputation quality higher than 0.4.

SNP-gene mapping: A number of 22,034 genes were downloaded from the UCSC (University of California Santa Cruz) genome browser (`http://genome.ucsc.edu/`) in RefFlat format (NCBI (National Center for Biotechnology Information) build 36/hg18). All genes that were assigned to more than one chromosome or had transcripts more than one mega base pairs long were excluded (71 genes) (as done by Segrè *et al.*, 2010). The genotyped and imputed SNPs were mapped to the genes according to their position. Genes that did not show any SNP assignments were excluded. Moreover, the analysis was reduced to autosomal chromosomes. That resulted in a number of 20,801 genes. In the following, the term gene is also used for the region in which SNPs were mapped to a gene, transcribed region of a gene with the flanking region of 110 kb upstream and 40 kb downstream. Such gene regions may be overlapping and some SNPs were mapped to several genes. The group of genes that were overlapping in at least one SNP is called *gene group*. For the presented data, the 20,801 genes lead to 2,319 gene groups. For the analysis of the LD effect a SNP-gene mapping was generated that included only independent SNPs that are approximately in linkage equilibrium with each other. LD pruning was performed with PLINK (Purcell *et al.*, 2007).

GWAS results: GWAS were calculated based on the described phenotype input and genotypes. SNPTEST v2.2.0 (Marchini *et al.*, 2007) was used for the calculation of the GWAS on the residuals for single phenotypes.

3.2. PSEA Results on Blood and Iron Traits

The selection of iron related phenotypes and blood traits enabled the usage of published results from large GWAS and meta-analyses for the evaluation of PSEA results. The results of enriched phenotype sets per gene were compared with published results of single phenotype GWAS. A gene is named *previously associated* gene for a trait if a SNP located in the gene (110 kb upstream to 40 kb downstream of the transcript) was published in a meta-analysis with a genome wide significant p-value (Benyamin *et al.*, 2009; Chambers *et al.*, 2009; Kullo *et al.*, 2010; Oexle *et al.*, 2011; Soranzo *et al.*, 2009;

Tanaka *et al.*, 2010). The previously associated genes named in the mentioned literature were presented in Figure 3.1 (details in appendix Table B.3). Enrichment of phenotype sets is neither limited to previously associated genes nor do all previously associated genes show enrichment of a phenotype set. However, if the enrichment of a phenotype set is significant for a gene that was previously published to be associated with elements of this phenotype set, it would be more likely a correct finding than the enrichment for a randomly selected gene.

Results of PSEA on Genotypes and Phenotypes per Person

For the results presented below 1,000 permutations were performed. As the step size of the p-values was determined by the number of permutations, the lowest p-value unequal to zero (0.001 for $1,000$ permutations) was taken as significance level. A phenotype set is called enriched for a gene if the p-value was lower than 0.001. As more than 50 tests were taken forward for replication, the replication p-value was set to 0.001.

For evaluation of PSEA results, the absolute count of genes that were identified by PSEA to be associated with at least one phenotype set was considered (shortly named: *identified genes*). The number of previously associated genes among these genes was used to assess the number of presumably true positive findings (Table 3.1). Furthermore, the number of corresponding gene groups was regarded as this number accounted for the overlapping gene definition. 16 % (43/272) of the genes that were identified were previously associated genes and therefore presumably true positive findings. Similar was observed for gene groups ($15/70 \approx 21$ % presumably true positive findings). One must be aware that the testing correction for the p-value criterion was possibly not strict enough. That fact could lead to a higher false positive rate. The replication step increased the percentage of presumably true positive findings to 67 %. In terms of gene groups it was even 100 %. But the increase in true positive rate was coincident with a decrease in the absolute number of identified previously associated genes. The detailed results of significantly enriched and replicated phenotype sets identified by PSEA are presented in Table 3.2.

Table 3.1.: **Number of PSEA results on blood and iron phenotypes.** The number of genes and gene groups that were identified with PSEA to be associated with at least one predefined phenotype set is presented. The column *previous* denotes how many of the identified genes/gene groups were previously known from published GWAS and meta-analyses.

		PSEA on genotypes and phenotypes				PSEA on GWAS results	
		KORA F4		replication in KORA F3		KORA F4	
		number of genes		number of genes		number of genes	
		identified	previous	identified	previous	identified	previous
$p < 0.001$	gene	272	43	52	35	58	22
	gene group	70	15	6	6	17	7

Identification of New Phenotype Sets

The significant and replicated results included one new phenotype set. The set, named *new set* in Table 3.2, consisted of iron, soluble transferrin receptor, transferrin saturation, MCH and MCV. It was significantly enriched in KORA F4 at *TMPRSS6* and the same set was identified and replicated in KORA F3. Altogether, 296 new phenotype sets were identified with PSEA in KORA F4, of which 162 were significantly enriched. Apart from the results presented in Table 3.1, many genes were identified to be associated only with new identified phenotype sets. The percentage of gene groups that include a previously associated gene among these identified gene groups is quite low (1.9 %, data not shown). Nevertheless, the replicated results showed that the identification of new phenotype sets can give valuable insight into phenotypic networks.

Comparison of PSEA with Single Phenotype GWAS

The findings of PSEA were compared with results of GWAS in KORA F4 on single phenotypes (Table 3.2). For three genes (two independent gene groups) none of the single phenotype GWAS on the blood and iron traits showed a significant result in KORA F4. Therefore, these genes would not have been found in GWAS in KORA F4 on single phenotypes but were identified by PSEA with enrichment of a phenotype set. For the remaining significant and replicated enrichments at least one phenotype per set had a genome wide significant p-value ($< 5 \cdot 10^{-8}$) in the single phenotype GWAS.

Table 3.2.: **Replicated PSEA results on blood and iron phenotypes.** Gene groups are separated by background colour. If the enrichment of a phenotype set was significantly enriched and replicated for more than one gene of a gene group, the gene with the highest NES is presented. The results for previously associated genes that were part of a gene group for which a significantly enriched and replicated phenotype set was identified are also reported in the table. Genes are marked with an asterisk if no SNP in the gene region has been published with an association of blood or iron traits before. The significance level for discovery and replication was < 0.001. For each gene the p-value and trait is given, that showed the lowest association p-value of all blood respectively iron traits and all SNPs in the gene in KORA F4 single phenotype GWAS. Gene names are indicated with bold letters if the minimal p-value of the KORA F4 GWAS was not genome wide significant (p-value $< 5 \cdot 10^{-8}$).

gene	pheno.-set	PSEA results discovery KORA F4 p-value	PSEA results replication KORA F3 p-value	GWAS results KORA F4 blood traits phenotype	blood traits p-value	iron tratis phenotype	iron tratis p-value
TMCC2	set_blood3	< 0.001	< 0.001	MPV	$4.10 \cdot 10^{-7}$	iron	0.0057
ARHGEF3	set_blood3	< 0.001	< 0.001	MPV	$3.03 \cdot 10^{-12}$	iron	0.0017
TF	set_iron1	< 0.001	< 0.001	PLT	0.0025	transferrin	$1.86 \cdot 10^{-41}$
TF	set_iron2	< 0.001	< 0.001	PLT	0.0025	transferrin	$1.86 \cdot 10^{-41}$
HFE	set_iron1	< 0.001	< 0.001	MCH	$1.41 \cdot 10^{-5}$	trans. sat.	$7.6 \cdot 10^{-9}$
HIST1H1C	set_iron1	< 0.001	< 0.001	MCH	$1.41 \cdot 10^{-5}$	trans. sat.	$7.6 \cdot 10^{-9}$
HFE	set_iron2	< 0.001	< 0.001	MCH	$1.41 \cdot 10^{-5}$	trans. sat.	$7.6 \cdot 10^{-9}$
HIST1H1A	set_iron2	< 0.001	< 0.001	MCH	$1.41 \cdot 10^{-5}$	trans. sat.	$7.6 \cdot 10^{-9}$
PCSK7	set_iron1	< 0.001	< 0.001	MCHC	0.0054	stfr	$2.25 \cdot 10^{-9}$
APOC3*	set_iron1	< 0.001	< 0.001	HCT	0.0017	stfr	$5.22 \cdot 10^{-10}$
PCSK7	set_iron2	< 0.001	< 0.001	MCHC	0.0054	stfr	$2.25 \cdot 10^{-9}$
APOC3*	set_iron2	< 0.001	< 0.001	HCT	0.0017	stfr	$5.22 \cdot 10^{-10}$
TMPRSS6	new set	< 0.001	< 0.001	MCV	$1.64 \cdot 10^{-4}$	stfr	$1.04 \cdot 10^{-6}$
TMPRSS6	set_iron1	< 0.001	< 0.001	MCV	$1.64 \cdot 10^{-4}$	stfr	$1.04 \cdot 10^{-6}$
C22orf33	set_iron1	< 0.001	< 0.001	MCV	$1.64 \cdot 10^{-4}$	stfr	$1.04 \cdot 10^{-6}$
TMPRSS6	set_iron2	< 0.001	< 0.001	MCV	$1.64 \cdot 10^{-4}$	stfr	$1.04 \cdot 10^{-6}$
C22orf33	set_iron2	< 0.001	< 0.001	MCV	$1.64 \cdot 10^{-4}$	stfr	$1.04 \cdot 10^{-6}$

stfr: soluble transferrin receptor; trans. sat.: transferrin saturation
The newly identified phenotype set *new set* consist of iron, soluble transferrin receptor, transferrin saturation, MCH and MCV.

Comparison with Published Results

PSEA revealed that the set of MPV and PLT is enriched for gene groups including *ARHGEF3* or *TMCC2*. These two genes were previously published in a large meta-analysis ($N > 13{,}500$) for an association with MPV. The alleles of the corresponding SNPs that increased the MPV were found to decrease the PLT. *ARHGEF3* was shown to be involved in the regulation of platelet counts and volume (intracellular signaling) (Soranzo *et al.*, 2009). The region including *TMCC2* was reported to be associated with MPV but no candidate gene was identified (Soranzo *et al.*, 2009). The results of PSEA showed that the set of all five iron traits (iron_set1) and the set of transferrin related traits (iron_set3: soluble transferrin receptor, transferrin and transferrin saturation) were significantly enriched and replicated for four gene groups including the genes *TF*, *HFE*, *TMPRSS6* and *PCSK7*. That corresponds to the associations that were published in large meta-analysis on different iron traits. *HFE* and *TF* were reported to be associated with transferrin (Benyamin *et al.*, 2009) and *HFE* and *TMPRSS6* with iron (Benyamin *et al.*, 2009; Tanaka *et al.*, 2010). Moreover, *HFE* as well as *TMPRSS6* were shown to have an indirect effect on the soluble transferrin receptor via the transferrin saturation. In contrast to that, the effect of *PCSK7* on soluble transferrin receptor was presumed to be more direct (Oexle *et al.*, 2011). The gene product of *TMPRSS6* is known to be involved in the regulation of levels of the peptide hormone hepcidin, which is an important master regulator of iron homeostasis in humans (Soranzo *et al.*, 2009). Additionally, *TMPRSS6* was identified in a large meta-analysis on blood traits to be associated with MCH, MCV and MCHC (Kullo *et al.*, 2010). The new identified phenotype set that was significantly enriched for *TMPRSS6* included five of the six mentioned traits that were previously reported for *TMPRSS6*. The findings of PSEA correspond to the published results from large meta-analyses.

Results for PSEA on GWAS Summary Statistics

PSEA was applied to KORA F4 GWAS results with 1,000 permutations of SNP based test statistics. In the analysis of PSEA on GWAS results 58 genes in 17 gene groups were

detected for which phenotype sets were significantly enriched (p-value < 0.001; Table 3.2). The comparison with results of PSEA on genotypes and phenotypes showed that all these genes, except two, were detected by PSEA on genotypes and phenotypes as well. Vice versa, the p-values of PSEA based on GWAS results for the replicated findings of PSEA on genotypes were considerably low (data not shown). In spite of the fact that the absolute number of identified genes is lower in PSEA based on GWAS results, it could be observed that the percentage of possibly true positive findings ($22/58 \approx 38\,\%$) was higher than for PSEA on genotypes. Despite the possible impacts on the permutation scheme, PSEA on GWAS results still found a valuable amount of presumably true findings.

3.3. Discussion of PSEA Results

This application of PSEA to iron and blood phenotypes demonstrated that valuable results can be achieved with PSEA. Findings from single phenotype analyses including large numbers of individuals could be reproduced with PSEA despite the lower sample size of the used KORA data.

PSEA Identifies more than Single Phenotype GWAS

PSEA identified several significant enrichments of the predefined phenotypes sets at different genes. The comparison of the results with published findings of large meta-analyses confirmed that a considerably high amount of the findings of PSEA were likely to be true positive findings. The application of PSEA to the given data identified several genes that were not genome wide significant in KORA F4 GWAS on the single phenotypes. In other words, PSEA could identify more loci that have an effect on multiple phenotypes than single phenotype GWAS. From phenotype sets that were significantly enriched for one or several genes information regarding the connection of phenotypes could be gained. In this application the observed connections were already known. However, in other situations the enrichment of phenotype sets might help to understand the unknown interdependencies of phenotypes.

Extension 1: Newly Identified Phenotype Sets

The application showed that the identification of new sets provided additional information. For *TMPRSS6* the information was gained that this gene has an effect on MCH and MCV apart from the effect on iron parameters. None of the predefined phenotype sets consisted of a mixture of iron and blood phenotypes. Without the possibility to identify new phenotype sets this interesting connection would have been missed.

The identification of new phenotype sets in PSEA used a fixed level of test statistics for single phenotype association. A data driven estimation of this threshold may be a point for further development.

Extension 2: Application of PSEA GWAS Results

The main benefit of the extension of PSEA to GWAS results is the applicability to situations where individual genotype and phenotype measurements are not available. In the application to GWAS results of blood and iron phenotypes in KORA F4 revealed similar results as for PSEA on genotypes and phenotypes per individuum.

It could be observed that the p-values of PSEA to GWAS results were a bit higher and with this not all findings of PSEA on individual genotypes and phenotypes could be found with the usage of GWAS results. This was mainly caused by the required modifications of the permutation scheme. As mentioned above there were two aspects that made the permutation scheme of PSEA on GWAS results less optimal than the permutation scheme of PSEA on genotypes and phenotypes: (A) possibly inflated test statistics distribution by association of phenotypes with other genes and (B) destroyed LD structure by SNP permutation. These two aspects were studied in this application to blood and iron phenotypes. For aspect (B) PSEA on GWAS results for a pruned list of SNPs was considered. With pruned GWAS results more genes were found than with PSEA on not pruned GWAS results. Apart from the increased absolute number of genes, also rose the number of identified genes that were previously published (Appendix Table B.4). Anyway, the percentage of previously published genes in all identified genes was even higher in PSEA on not pruned GWAS results. The destroyed LD structure

reduced the absolute number of identified genes but PSEA on GWAS results still led to interesting results. The aspect (A) is highly dependent on the phenotypes under consideration. For the present data, the percentage of genes that were identified by PSEA on genotypes and with PSEA on GWAS results (pruned data) decreased with the number of associated genes and strength of association of the phenotypes in the phenotype set under consideration in the single phenotype GWAS (Appendix Table B.5 and Appendix Figure B.1)). But in many cases especially phenotypes with no strong effect on a single phenotype would be interesting to test for phenotype set enrichment. For those phenotypes the effect of (A) would be small.
The results showed that even with not pruned data and inflated test statistic distributions PSEA on GWAS results can identify interesting gene phenotype set relations.

Computer Intensity

As mentioned above, the algorithm of PSEA was implemented with the programming language C with MPI parallelization. The program was executed on the Edinburgh Parallel Computing Centre (EPCC) supercomputing platform HECToR (High End Computing Terascale Resources) phase 2a (12,288-processor Cray XT4) within a project of DEISA (Distributed European Infrastructure for Supercomputing Applications). On 100 nodes, each with four processes, a genome wide run with 2.8 million SNPs, 28,000 genes, 14 phenotypes for 1,814 individuals and six phenotype sets with 1,000 permutations took around 105 minutes per cohort. In terms of computer intensity of the permutation strategies PSEA on GWAS results is clearly less demanding than PSEA on individual genotypes and phenotypes. A genome wide run for the mentioned phenotypes and phenotype sets on 2.18 million SNPs (only SNPs with good quality) took approximately 38 hours on one core of an Intel core i7 975 extreme 3.33 GHz, 24 GB RAM Linux computer.

This application demonstrated, that PSEA can detect additional loci that were not found with single phenotype GWAS on the same data. In situations were the relation of the phenotypes is unknown, for example in metabolomics data, PSEA could also improve the knowledge on the network of phenotypes.

4. Application to Metabolomics Data – Predefined Phenotype Sets

This chapter describes the application of the general algorithm of PSEA to two different panels of metabolomics data. It is shown that PSEA can improve the knowledge on the complex and high dimensional network of intermediate phenotypes. Furthermore, it demonstrates the ability of PSEA to cope with a large number of phenotypes. At first metabolomics data is described. A special aspect of this application is the generation of metabolite sets with the Gaussian graphical modeling (GGM) that uses partial correlation between the metabolites. Data of KORA F4 was used for discovery and replication was performed in data from the TwinsUK study. The results of PSEA for both metabolite panels are presented and selected results are discussed biologically. A more methodological view on the application is given in the end of this chapter.

4.1. Metabolomics Data

For the analysis data from KORA F4 (described in section 3.1) was used as discovery cohort and data from the TwinsUK cohort as replication. The TwinsUK study is a British adult twin-registry. The participants were recruited from the general population through national media campaigns in the United Kingdom and were shown to be comparable to age-matched population singletons in terms of disease related and lifestyle characteristics (Andrew *et al.*, 2001). Measurements of metabolomics data with two different technologies was performed in both cohorts.

Biocrates Metabolites

A panel of 163 metabolites was measured for individuals of KORA F4 using electro spray ionization tandem mass spectrometry with the AbsoluteIDQ kit (BIOCRATES Life Sciences AG). Details of the measurement methods were described in previous publications (Gieger *et al.*, 2008; Illig *et al.*, 2010). Metabolites that were not stable (experimental variance $> 25\,\%$) in repeated measurements of some control samples were excluded. Moreover, metabolites with more than $10\,\%$ missing values were removed from the analyses. Per metabolite extreme values (± 5 standard deviations from mean value) were identified. Individuals that showed extreme values for more than three independent metabolites (correlation $< 70\,\%$) were excluded. Extreme values for remaining individuals were set to missing. Missing values were imputed with the MICE algorithm (Multivariate Imputation by Chained Equations; `http://cran.r-project.org/web/packages/mice/index.html`) that was implemented in R (`http://www.r-project.org/`). After quality control 151 metabolites remained for further analyses. These 151 metabolites can be grouped in ten metabolite classes including amino acids (14), hexoses (1), different carnitines (free carnitine (1), acylcarnitines (22), hydroxy- and dicarboxyacylcarnitines (12)), sphingomyelins (9) and hydroxysphingomyelins (5) and different forms of phosphatidylcholines (diacyl-phosphatidylcholines (36), acyl-alkyl-phosphatidylcholines (38), lyso-phosphatidiylcholines (13)). A full list of all metabolites is given in the appendix Table C.1. For 1,809 individuals in KORA F4 Biocrates metabolites and genome wide genotypes were available.

For replication data from the TwinsUK cohort was used that had measurements for metabolites with the same AbsoluteIDQ kit (BIOCRATES Life Sciences AG). The metabolites underwent the same quality control as described for KORA F4. All 151 metabolites passed the quality control in TwinsUK as well. 1,173 individuals with genotypes and valid Biocrates metabolites measurements were used for further analysis. After reduction to unrelated individuals 843 individuals remained in the analysis.

Metabolon Metabolites

A different set of 295 metabolites were measured for individuals of KORA F4 with a technique supplied by Metabolon. It used ultrahigh-performance liquid-phase chromatography and gas-chromatography separation with tandem mass spectrometry (Evans *et al.*, 2009; Ohta *et al.*, 2009). The measurement method was described in detail in a previous publication (Suhre *et al.*, 2011). 102 metabolites had more than 10 % missing values and were excluded from the analyses. Missing values for the remaining metabolites were imputed with MICE (see description for Biocrates metabolites). The remaining 193 metabolites spanned different super pathways including amino acids (52), carbohydrates (10), cofactors and vitamins (7), energy (3) and lipid (90) pathway relevant compounds, nucleotides (9), peptides (11) and xenobiotics (11). A full list of all 193 metabolites together with additional information about the pathways they belong to is given in the Appendix Table C.2. In total 1,768 individuals with valid Metabolon metabolites measurements and genotypes could be used for further analysis.

The same technology was used to measure metabolites in TwinsUK data. Only metabolites that passed quality control in KORA F4 were regarded. Individuals with more than 50 % missing values were excluded. That led to 1,052 individuals with a maximal missing rate of 21.16 %. Four metabolites that were present in KORA F4 had less than 300 valid measurements in TwinsUK data. According to Suhre *et al.* (2011), 300 is the critical limit of non-missing values to avoid false positive findings due to small sample size. Therefore, these four metabolites were excluded from further analysis. In the remaining 189 metabolites the maximal missing rate per metabolite was 65.59 %, which is equivalent to 362 valid measurements. To assure that most metabolite sets could be analyzed in the replication, no further exclusion criteria for metabolites were applied. No imputation of missing data was performed. After reduction to unrelated and genotyped individuals 705 individuals remained in the analysis.

Definition of Metabolite Sets – Gaussian Graphical Modeling (GGM)

For the definition of phenotype sets for PSEA GGM was applied, which used the conditional dependence between the variables (Krumsiek *et al.*, 2011). The models were estimated based on the partial correlation coefficient of each pair of metabolites, which is the pairwise Pearson correlation coefficient conditioned against the correlation with all other metabolites in the analyses. This approach was shown to be a valuable tool for identification of metabolite networks, which is able to distinguish direct form indirect associations (Krumsiek *et al.*, 2011). Another advantage is that this estimation of metabolite sets is independent from further information like availability of database information. The analysis strategy was applied to the panel of all metabolite measurements in KORA F4 that passed quality control and to all individuals with metabolite measurements and genotypes. Metabolon and Biocrates metabolites were analyzed separately. Two partial correlation coefficient cutoff levels (0.3 and 0.45) were used. Both cutoffs gave several metabolite sets. The sets were not overlapping for each cutoff, but sets gained from the higher cutoff level were likely to be subsets of the sets gained with the lower cutoff level. This could be observed in the metabolite sets.

For Biocrates metabolites this approach led to 40 metabolite sets (presented in Figure 4.1). Two metabolite sets were defined with both cutoff levels. That means 38 different metabolite sets were found. The size of the metabolite sets ranged from 2 to 27. One can see from Figure 4.1 that the sets often include metabolites of related metabolite classes. For example six sets include carnitines exclusively. Different forms of phosphatidylcholines were found to be in a set with each other or with sphingomyelins. 116 metabolites were part of at least one phenotype class, that means for 35 metabolites there was no metabolite with a pairwise correlation higher than 0.3. A detailed list of metabolites in each metabolite set is given in Appendix Table C.3.

The metabolite sets for Metabolon metabolites presented in Figure 4.2 were determined in the same way. For this data 50 different metabolite groups were identified (13 metabolite sets were detected by both cutoff values). The size of the metabolite sets ranged between 2 and 20. The metabolite sets for Metabolon metabolites also included often related

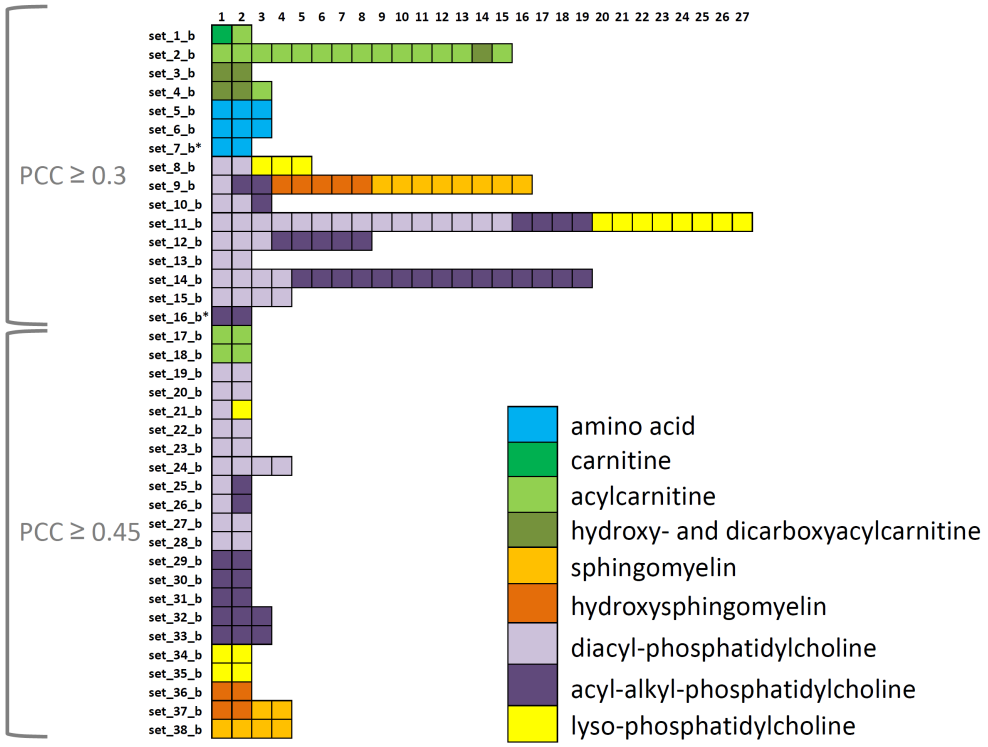


Figure 4.1.: **GGM-defined Biocrates metabolite sets.** Each metabolite of a set is represented by a box coloured according to its metabolite class. The sets that were identified with both partial correlation coefficient (PPC) cutoffs are marked with a star.

metabolites (see Figure 4.2). For example many sets contained only metabolites of lipid pathways, two sets consisted of a combination of lipid and energy pathway metabolites (detailed list of metabolites in each metabolite set is given in Appendix Table C.4). All in all, 115 metabolites were part of at least one metabolite set. In the metabolite measurements of the TwinsUK, which was used for replication, two of these 115 metabolites were missing. Therefore, for the replication two metabolite sets had to be modified, as they include at least one of these missing metabolites. One set was made up only from these two metabolites and had to be excluded from the replication. The other set included several other metabolites and could be tested in a modified form.

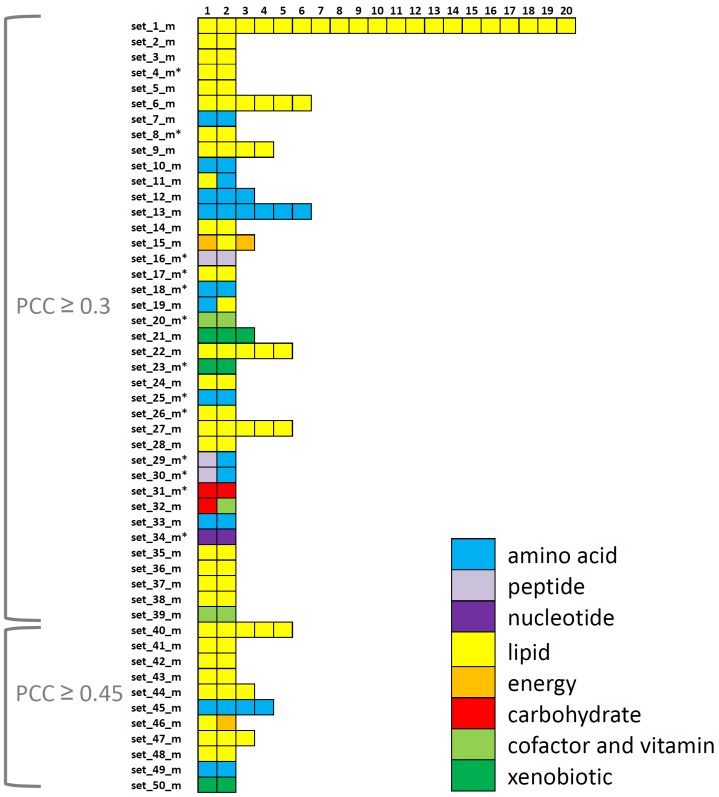


Figure 4.2.: **GGM-defined Metabolon metabolite sets.** Each metabolite of a set is represented by a box coloured according to its metabolite super pathway. The sets that were identified with both partial correlation coefficient (PPC) cutoffs are marked with a star.

Genotypic and Phenotypic input

For KORA F4 the same genotyping and imputation was used as described in section 3.1. The genotyping of the replication cohort TwinsUK was performed with a combination of different Illumina arrays (HumanHap300, HumanHap510Q, 1M-Duo and 1.2MDuo 1M). For each set Illuminus calling algorithm was used (Duo arrays pooled for calling). After sample and SNP-wise filtering the data was pooled and imputed using Impute v2 (Marchini and Howie, 2010) (reference HapMap 2 and the combined HumanHap610k and 1M reduced to 610k SNP content). More details on genotyping were given in previous publications (e.g. Soranzo *et al.*, 2009; Suhre *et al.*, 2011). The analysis was restricted

to unrelated individuals. SNPs were filtered for minor allele frequency higher than 5 %, call rate higher than 95 % and imputation quality higher than 0.4.

For both Metabolon and Biocrates metabolites outliers that differed more than five standard deviation from the mean were excluded. The residuals of log-transformed metabolites with adjustment for sex and age were calculated and taken as phenotypic input for PSEA. For Biocrates metabolites additional adjustment for an internal batch variable accounting for possible measurement differences was applied. It could be seen that after log-transformation most (146) Biocrates metabolites were closer to the normal distribution than untransformed metabolite concentrations. For Metabolon metabolites the same was previously shown with log10-transformation (Suhre *et al.*, 2011). For simplicity per panel the same transformation was applied to all metabolites.

For PSEA an approach was applied that increased the number of permutations stepwise. It started with 100 permutations calculated for all genes. Genes, for which at least one phenotype set was enriched with a p-value ≤ 0.03, were taken forward for the next step of 1,000 permutations. In the third step, 10,000 permutations were performed for genes with a p-value ≤ 0.003. The lowest possible p-value 10^{-4} was taken as significance level. Replication of all significant enrichments was calculated in TwinsUK with 10,000 permutations (see Appendix Figure C.1).

4.2. PSEA Results on Biocrates Metabolites

PSEA on KORA F4 Biocrates metabolites found 35 phenotype sets significantly enriched for at least one gene (p-value $\leq 10^{-4}$). 163 genes were identified, which belonged to 61 different gene groups. All in all, 354 phenotype set gene combinations were significant including 35 different phenotype sets and 163 genes in 61 gene groups. The 163 genes were taken forward for replication in TwinsUK data. After correction for 61 independent gene groups, 129 of the 354 significant enrichments were replicated (significance level: $0.05/61 \approx 0.00082$). These 129 enrichments were identified at six different gene groups. The replicated results are presented in Table 4.1.

Table 4.1.: **Replicated PSEA results on GGM-defined Biocrates metabolite sets.** For the presented genes the given phenotype sets showed an enrichment with a p-value $< 10^{-4}$ in PSEA on KORA F4 and a p-value $< 8.2 \cdot 10^{-4}$ on TwinsUK data. Gene groups are separated with background colour.

	start pos.	stop pos.	phenotype sets (*set_ *_ b*)
chr. 1	75852750	76280263	17
genes:	ACADM, RABGGTB, SNORD45A, SNORD45B, SNORD45C, MSH4, ASB17		
chr. 2	210942386	211291577	5
genes:	CPS1		
chr. 4	159702854	160087236	17
genes:	ETFDH , C4orf46, FNIP2, PPID		
chr. 10	61042363	61248406	1
genes:	SLC16A9		
chr. 11	61094821	61430694	11, 14, 23, 24, 30, 31, 32, 33, 35, 8
genes:	C11orf9, FEN1, DKFZP434K028, DAGLA, FADS2,		
chr. 11	61273486	61527475	11, 14, 23, 24, 30, 31, 32, 33, 35
genes:	C11orf10, MIR611, FADS1, MIR1908, FADS3, BEST1		
chr. 11	61381461	61551489	23, 24, 31, 35
genes:	RAB3IL1		
chr. 14	63189959	63799513	36
genes:	SGPP1, SYNE2		
chr. 14	66884273	67165389	16, 30
genes:	PLEK2, PLEKHH1, TMEM229B		

Enrichment of Amino Acid Sets

At the *CPS1* gene an enrichment of the *set_ 5_ b* was replicated. This set consisted of three amino acids (glycine, serine and threonine).

Enrichment of Lipid Pathway Related Sets

Two non overlapping sets (*set_ 16_ b*, *set_ 30_ b*), which included both two acyl-alkyl-phosphatidylcholines, were significantly enriched for three genes (*PLEK2*, *PLEKHH1*, *TMEM229B*) in a region on chromosome 14. On chromosome eleven there was a region including the FADS cluster (*FADS1*, *FADS2*, *FADS3*), in which different phenotype sets showed enrichment at twelve genes. These phenotype sets included diacyl-phosphatidylcholines, acyl-alkyl-phosphatidylcholines and lyso-phosphatidylcholines. Four sets (*set_ *_ b*: 23, 24, 31, 35) were enriched for all twelve genes. For the other genes respectively six (*set_ *_ b*: 11, 14, 30, 32, 33) or seven phenotype sets (additionally *set_ 8_ b*) were significantly enriched and replicated. It could be observed that the identified sets were overlapping especially with *set_ 11_ b* and *set_ 14_ b*, which were the largest sets identified for Biocrates metabolites. At last the phenotype set *set_ 36_ b* that consisted

of two hydroxysphingomyelins (hydroxysphingomyelin C14:1 and C16:1) was found to be enriched for *SGPP1* and *SYNE2* on chromosome 14.

Enrichment of Carnitine Sets

The phenotype set *set_17_b*, which included two acylcarnitines (decanoylcarnitine, octanoylcarnitine), was enriched and replicated for seven genes, including *ACADM* in a gene group on the first chromosome. The enrichment of the same phenotype set was replicated for *ETFDH* and other genes of a gene group on chromosome four. Another phenotype set (*set_1_b*) that consisted of free carnitine and propionylcarnitine was enriched and replicated for *SLC16A9*.

4.3. PSEA Results on Metabolon Metabolites

PSEA on Metabolon metabolite sets in KORA F4 revealed 344 significant enrichments of phenotype sets. 35 metabolite sets had at least one significant enrichment. These enrichments were located at 252 different genes, which could be grouped into 58 independent gene groups. In the replication step these 344 significant enrichments were analyzed on TwinsUK data. The replication significance level was corrected for the 58 not overlapping gene regions (replication significance level: $0.05/58 = 8.6 \cdot 10^{-4}$). 183 enrichments at 66 genes (8 gene groups) and 10 phenotype sets could be replicated in TwinsUK data. The detailed results are presented in Table 4.2.

Results of Amino Acids Sets

A set of creatine and pyroglutamine (*set_25_m*) as well as a set of glycine and serine (*set_33_m*) were enriched for *CPS1*. In a larger gene region on chromosome twelve a set that consisted of a peptide (γ-glutamyl-glutamine) and a amino acid (glutamin) (*set_29_m*) was enriched and replicated. Among others *GLS2* was identified.

Results of Carbohydrate Sets

Only one metabolite set of the carbohydrate pathway was significantly enriched and replicated. At a region on chromosome two including *GCKR* and four other genes the metabolite set of glucose and mannose (*set_31_m*) was enriched.

Table 4.2.: **Replicated PSEA results on GGM-defined Metabolon metabolite sets.** For the presented genes the given phenotype sets showed an enrichment with a p-value $< 10^{-4}$ in PSEA on KORA F4 and a p-value $< 8.6 \cdot 10^{-4}$ on TwinsUK data. Gene groups are separated with background colour.

	start pos.	stop pos.	phenotype sets (*set_ *_ m*)
chr. 1	75402838	76907932	47, 9
genes:	SLC44A5, ACADM, RABGGTB, SNORD45A, SNORD45B, SNORD45C, MSH4, ASB17, ST6GALNAC3		
chr. 2	27464973	27766306	31
genes:	GCKR, FNDC4, C2orf16, ZNF512, GPN1		
chr. 2	210942386	211291577	25, 33
genes:	CPS1		
chr. 2	234009939	234427219	20
genes:	USP40, UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1, DNAJB3		
chr. 4	159702854	159973810	47, 9
genes:	ETFDH, C4orf46, PPID		
chr. 4	159808181	160087236	9
genes:	FNIP2		
chr. 7	98700955	99508811	17
genes:	ARPC1B, BUD31, CPSF4, PDAP1, ZNF789, PTCD1, ZKSCAN5, ATP5J2, ZNF655, ZNF394, FAM200A, ZNF498 CYP3A5, CYP3A7, CYP3A4, CYP3A43, OR2AE1, TRIM4, GJC3, ZKSCAN1		
chr. 11	61166755	61430694	40, 41
genes:	C11orf9, FEN1, DKFZP434K028, FADS2		
chr. 11	61273486	61525549	40
genes:	C11orf10, MIR611, FADS1, MIR1908, FADS3		
chr. 12	54985890	55309551	29
genes:	STAT2, APOF, SPRYD4, TIMELESS, MIP, RBMS2, GLS2		

Results of Cofactors and Vitamins

On chromosome 10 there were several genes of the UGT1A UDP glucuronosyltransferase 1 family (and three other genes) that were identified for a significant enrichment of *set_ 20_ m* consiting of bilirubin(E;E) and bilirubin(Z;Z).

Results of Carnitine Sets

Significant enrichments of two phenotype sets (*set_ 9_ m* and *set_ 47_ m*) were replicated for nine genes of one region on the first chromosome including *ACADM*. The same set was enriched and replicated for *ETFDH* and two other genes (region on chromosome 4). In the same region there was one gene (*FNIP2*) for which only the enrichment with *set_ 9_ m* could be replicated. The set *set_ 47_ m* consisted of three carnitines: carnitine, hexanoylcarnitine and octanoylcarnitine. These three metabolites were also elements of *set_ 9_ m* besides a fourth carnitine (2-tetradecenoyl carnitine).

Results of Lipid Pathway Related Sets

In a region on chromosome seven several genes including *CYP3A5* were found with an enrichment of *set_17_m* that included two metabolites of the lipid/sterol/steroid pathway (androsterone, epiandrosterone). The region including the *FADS*-cluster showed enrichment with *set_40_m* and *set_41_m* or only with *set_40_m*. Both sets were not overlapping and included different glycerolipids, lysolipids or fatty acids.

4.4. Discussion of PSEA Results on Predefined Phenotype Sets

PSEA on metabolomics data with usage of the predefined sets derived from the GGM revealed several loci that showed enrichment with one or more metabolite sets. As some metabolites were measured by both panels or closely related metabolites were covered, it was not surprising that three gene regions were identified for both metabolite panels. These gene regions include *ACADM*, *CPS1* and several genes surrounding the *FADS* cluster. Additionally, four gene regions were identified only with Biocrates data and five with Metabolon data only.

The enriched phenotype sets carried valuable information on the underlying processes, which can be seen exemplarily for the enrichment at *ACADM*:
The enriched sets for *ACADM* consisted of decanoylcarnitine and octanoylcarnitine for Biocrates data and of carnitine, hexanoyl-, octanoyl- and 2-tetradecenoyl carnitine for Metabolon data. The gene product of *ACADM*, the medium chain acyl-coenzyme A dehydrogenase, catalyzes the first step of β-oxidation. The medium chain acyl-coenzyme A dehydrogenase supports the generation of the acyl-coenzyme A with carnitine from medium chain acylcarnitine. Therefore acylcarnitines with medium chain length $(4-12)$ are known to be substrates of the gene product of *ACADM* (Nichols *et al.*, 2008). This biological connection was reflected in the results of PSEA.
In GWAS *ACADM* was found to be associated with concentrations of all phenotype set elements except for carnitine and 2-tetradecenoyl carnitine but these were part of

significantly associated ratios.

Further comparison of PSEA findings with GWAS results showed, that all identified genetic loci were previously associated in single phenotype analyses on metabolite concentrations or ratios on the same data (Illig *et al.*, 2010; Suhre *et al.*, 2011). It could be observed that the identified enriched phenotype sets corresponded to the published single phenotype association. The gene function was discussed in the related publications (Gieger *et al.*, 2008; Illig *et al.*, 2010; Suhre *et al.*, 2011).

On the one hand, the application showed that PSEA is able to identify loci that are associated with sets of metabolites. The phenotype sets that were enriched for a specific gene carry information about the underlying processes as seen for the example of *ACADM*. On the other hand, new genes could not be identified.

One reason for this could be that loci identified for metabolites often mirror the substrate affinity of the coded enzymes. Therefore, the strength of association is very high, e.g. in GWAS on Biocrates metabolites p-values are observed in the size of $3.5 \cdot 10^{-78}$ (C4, *ACADS*) for concentrations and in the size of $6.5 \cdot 10^{-179}$ (PC.aa.C36.3/PC.aa.C.36.4, *FADS*) for ratios. On Metabolon metabolites the minimal p-values are even lower: $5.4 \cdot 10^{-252}$ (N-acetylornithine, *NAT8*) for concentrations and $< 4.4 \cdot 10^{-305}$ (Butyrylcarnitine/propionylcarnitine, *ACADS*) for ratios. Such stong associations could easily be detected in single phenotype GWAS.

Another reason might be the determination of predefined metabolite sets with GGM. It could be observed that the GGM-defined phenotype sets represent sensible and important biological processes. This is confirmed as the found enrichments were conform with the GWAS results. But apparently, the hypotheses proposed by GGM did not led to new identified loci. As any definition of phenotype sets lead to a limitation of hypotheses, this is not a special problem of the usage of GGM. In fact the limitation is even necessary to reduce the computing time of PSEA. One opportunity to enlarge the hypotheses in addition to the predefined phenotype sets is the internal identification of new phenotype sets, which is one extension of PSEA.

5. Application to Metabolomics Data – Identification of New Phenotype Sets

The internal identification of new phenotype sets, made possible by the first extension of PSEA, expands the hypotheses of predefined phenotype sets with promising new identified sets. In this chapter the application of this extension on Biocrates and Metabolon metabolite panels is presented. Selected results will be discussed biologically. Comparisons between the approach of predefined phenotype sets and the internal new identification of phenotype sets are discussed, with respect to the potential of the latter to unravel additional information.

5.1. Identification of New Metabolite Sets

The extension of PSEA to detect new phenotype sets was used as described in section 2.2. All metabolites for which a gene had an association p-value lower than $5 \cdot 10^{-4}$ were grouped to a new phenotype set for this genes and analyzed for enrichment. The same metabolites and genotypes of KORA F4 and TwinsUK with the same quality control were used as described in section 4.1.

For a large number of genes new phenotype sets (11,607 Biocrates, 14,993 Metabolon) were identified. As most of the new identified phenotype sets (shortly named: *new phenotype sets*) had a low p-value (≤ 0.003) after 1,000 permutations (10,850 Biocrates, 14,781 Metabolon), an *intermediate* replication step was performed, to reduce the computational effort. That means, all genes at which a phenotype set, newly identified in KORA F4,

had a p-value ≤ 0.003 after 1,000 permutations were replicated in TwinsUK with 1,000 permutations. Only those genes at which a new phenotype set gained a p-value ≤ 0.003 in PSEA on TwinsUK with 1,000 permutations (intermediate replication) were analyzed in KORA F4 with 10,000 permutations. For all significant findings of PSEA in KORA F4 with 10,000 permutations (p-value < 0.0001) replication in TwinsUK with 10,000 was performed (see Appendix Figure D.1).

5.2. PSEA Results on Biocrates Metabolites

A number of 62 genes showed a significant enrichment of new phenotype sets (p-value $<$ 0.0001). These genes belonged to 22 independent gene groups. A significant enrichment of a phenotype set is counted only once per gene group even if a new phenotype set was identified and significantly enriched for several genes of a gene group. That led to a number of 45 significant enrichments. After correcting the replication p-value for 45 independent enrichments ($0.05/45 \approx 0.001$), 46 findings could be replicated including 15 gene groups and 33 new phenotype sets. All new identified, significantly enriched and replicated sets are presented in Figure 5.1 and Appendix Table D.1.

In general, the new identified phenotypes sets were larger than the predefined phenotype sets. 15 of the significantly enriched and replicated new metabolite sets included more than 20 metabolites. In contrast to that, only one of the predefined phenotype sets included more than 20 metabolites. Moreover, one could observe that the new phenotype sets included more metabolites of different metabolite classes.

Six gene groups that were associated with a new set showed a significant and replicated enrichment with a predefined set as well (see Figure 5.1). The new phenotype sets for these genes included at least some elements of the enriched predefined sets, but often more metabolites than this. For example, the new set at *ETFDH* consisted of two acylcarnitines (C8, C10) and one sphingomyelin (SM.C26.0). The predefined set that was enriched for *ETFDH* is a subset of these metabolites including the two acylcarnitines. A more extreme example is the enrichment of *SLC16A9*. The new phenotype set included

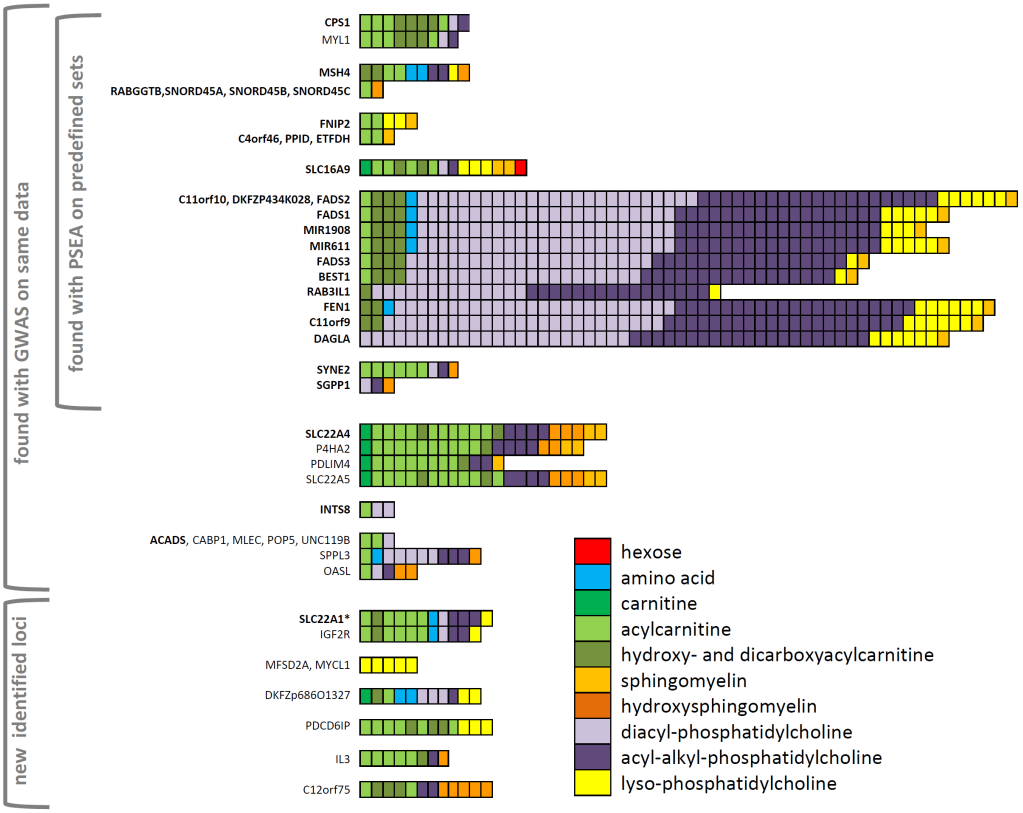


Figure 5.1.: **New identified sets for Biocrates metabolites.** All new metabolite sets that were significantly enriched in KORA F4 and replicated in TwinsUK are presented along with the genes at which they were identified. Each metabolite of a set is represented by a box coloured according to its metabolite class. The brackets on the left side state if at least one gene of the gene group was identified with PSEA for an enrichment of a GGM-defined set on the same data, with GWAS on the same data or if the locus is new. The genes that were previously identified with the named analyses are printed in bold letters. *SLC22A1* was found in GWAS to be associated with one metabolite of the Metabolon data but not with Biocrates, therefore it is marked with a star.

15 metabolites of eight different metabolite classes. The predefined phenotype set that was significantly enriched and replicated for this gene (C0, C3) is a subset of the new phenotype set. Similar observations could be made for the gene region including *CPS1*. The predefined phenotype set enriched for *CPS1* consisted of three amino acids. Besides two of these amino acids, different forms of carnitines and phosphatidylcholines were included in the new phenotype sets. The largest significantly enriched and replicated new phenotype sets were identified at the *FADS* cluster. For *FADS2* a phenotype set of 58 metabolites was replicated. Lipid related metabolites made up the major fraction of the identified sets at this gene group. At *SYNE2/SGPP1* the identified phenotype sets included one hydroxysphingomyelin, two phosphatidylcholines and at *SYNE2* some acylcarnitines. Only one hydroxyspingomyelin was part of both the new phenotype sets and the significantly enriched predefined phenotype sets at these genes.

The remaining nine gene groups were not found with PSEA on predefined metabolite sets neither on Biocrates nor on Metabolon data. These gene groups include the additional information gained with internal identification of new phenotype sets with PSEA.

5.3. PSEA Results on Metabolon Metabolites

242 genes of 63 independent gene groups were identified to have a significant enrichment with a new phenotype set in KORA F4. 39 of these significant findings (19 phenotype sets) could not be analyzed for replication as less than two metabolites of each set were available in TwinsUK data (results presented in Appendix Table D.3). The significance level of the replication step had to be corrected for the number of significant findings. Each new phenotype set in the remaining 203 significant enrichments was counted only once per gene group. These led to a number of 107 significant enrichments. With this correction ($0.05/107 \approx 0.0005$) 131 enrichments could be replicated in TwinsUK data. This included 67 different phenotype sets and 131 genes at 23 gene groups. The results are presented in Figure 5.2 and in Appendix Table D.2.

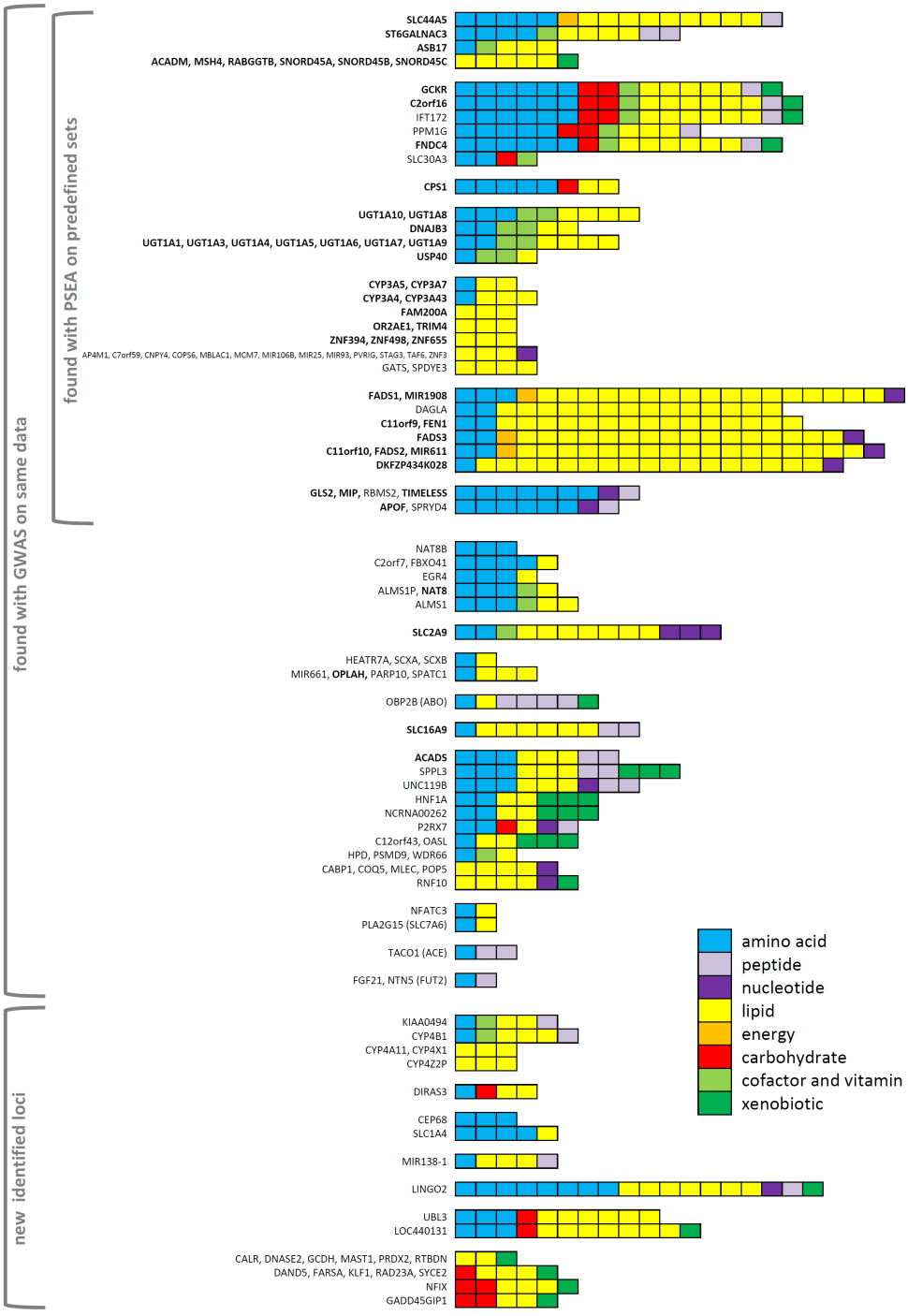
Same observations concerning size and diversity of the new phenotype sets could be made

as described above for new sets on Biocrates data.

Several genes of seven gene groups were significantly enriched and replicated with predefined phenotype sets (indicated in Figure 5.2). At some gene groups the metabolites of the new phenotype sets belonged to the same metabolite classes as the metabolites of the GGM-defined sets. For example the new phenotype sets of *GLS2, MIP, TIMELESS, APOF* and *SPRYD4* included the amino acid and peptide of the predefined set besides additional amino acids and nucleotides. For the gene group including the *FADS* gene, which showed enrichment for predefined sets of metabolites related to the lipid metabolism, the new phenotype sets were observed to include mainly metabolites of the lipid metabolism as well. The same was true for genes of a gene region including *CYP3A4*. For the other four gene groups, that showed significant enrichment with predefined and new phenotype sets, the new phenotype sets included metabolites of more diverse pathways. At *ACADM* and surrounding genes several carnitines were included in

---

Figure 5.2. *(on the next page)*: **New identified sets for Metabolon metabolites.** All new identified metabolite sets that were significantly enriched in KORA F4 and replicated in TwinsUK are presented along with the genes at which they were identified. Each metabolite of a set is represented by a box coloured according to its superpathway. Not overlapping gene groups are separated with horizontal spaces. The brackets on the left side state if at least one gene of the gene group was identified with PSEA on a GGM-defined set on the same data, with GWAS on the same data or if the locus is new. The genes that were previously identified with the named analyses are printed in bold letters. If not the gene which is presented here but another gene of the gene group was previously identified, the known gene is given in brackets.


found with GWAS on same data
found with PSEA on predefined sets
new identified loci
SLC44A5
ST6GALNAC3
ASB17
ACADM, MSH4, RABGGTB, SNORD45A, SNORD45B, SNORD45C
GCKR
C2orf16
IFT172
PPM1G
FNDC4
SLC30A3
CPS1
UGT1A10, UGT1A8
DNAJB3
UGT1A1, UGT1A3, UGT1A4, UGT1A5, UGT1A6, UGT1A7, UGT1A9
USP40
CYP3A5, CYP3A7
CYP3A4, CYP3A43
FAM200A
OR2AE1, TRIM4
ZNF394, ZNF498, ZNF655
AP4M1, C7orf59, CNPY4, COPS6, MBLAC1, MCM7, MIR106B, MIR25, MIR93, PVRIG, STAG3, TAF6, ZNF3
GATS, SPDYE3
FADS1, MIR1908
DAGLA
C11orf9, FEN1
FADS3
C11orf10, FADS2, MIR611
DKFZP434K028
GLS2, MIP, RBMS2, TIMELESS
APOF, SPRYD4
NAT8B
C2orf7, FBXO41
EGR4
ALMS1P, NAT8
ALMS1
SLC2A9
HEATR7A, SCXA, SCXB
MIR661, OPLAH, PARP10, SPATC1
OBP2B (ABO)
SLC16A9
ACADS
SPPL3
UNC119B
HNF1A
NCRNA00262
P2RX7
C12orf43, OASL
HPD, PSMD9, WDR66
CABP1, COQ5, MLEC, POP5
RNF10
NFATC3
PLA2G15 (SLC7A6)
TACO1 (ACE)
FGF21, NTN5 (FUT2)
KIAA0494
CYP4B1
CYP4A11, CYP4X1
CYP4Z2P
DIRAS3
CEP68
SLC1A4
MIR138-1
LINGO2
UBL3
LOC440131
CALR, DNASE2, GCDH, MAST1, PRDX2, RTBDN
DAND5, FARSA, KLF1, RAD23A, SYCE2
NFIX
GADD45GIP1
amino acid
peptide
nucleotide
lipid
energy
carbohydrate
cofactor and vitamin
xenobiotic

the new phenotype sets. This corresponds to the enriched predefined phenotype sets at this gene group. But apart from that also some fatty acids and metabolites of the amino acid super pathway were part of the new phenotype sets. Genes at the gene region of *GCKR* were identified with the predefined phenotype set of glucose and mannose. The new phenotype sets at these gene included up to 17 different metabolites. They consisted of the carbohydrates lactate, mannose and threitol (nucleotide sugar) but also various metabolites of the glycerolipid metabolism and amino acids of the glutamate metabolism. Both predefined phenotype sets that were significantly enriched for *CPS1* included two amino acids. These four amino acids were part of the new phenotype sets at *CPS1* along with two more amino acids, two fatty acids and the amino sugar erythronate. For the region around the *UGT1A* gene the new phenotype sets included, in addition to the two isomers of billirubin that made up the predefined set, some metabolites of the lipid and amino acid metabolism. The remaining 16 gene groups were new, as they were not identified with any predefined phenotype set.

5.4. Discussion of the PSEA Results on New Identified Phenotype Sets

PSEA with the extension for identification of new phenotype sets identified on Biocrates data 15 gene groups and on Metabolon data 23 gene groups, that showed enrichment of a new metabolite set. As mentioned above, six respectively seven gene groups were found in the analysis of predefined phenotype sets as well. The comparison with GWAS on the same data revealed that additional three gene groups of the Biocrates results and nine gene groups of the Metabolon results are significantly associated with at least one metabolite concentration or ratio (Illig *et al.*, 2010; Suhre *et al.*, 2011) (see Figures 5.1 and 5.2). One of the new loci (*SLC22A1*) identified on Biocrates data was found with a significant association in GWAS on Metabolon data. In total, five new loci on Biocrates data and seven on Metabolon data were not found in the GWAS on the metabolite data sets. The elements of the identified and enriched metabolite sets enabled insight into the

underlying processes of genes, their products and various metabolites. Three genetic loci that are not associated in a single phenotype GWAS on the metabolite data but were found with PSEA are discussed here exemplarily. The remaining nine new genetic loci are discussed in the Appendix section D.2. Interpretation of genes known from GWAS can be found in the corresponding publications (Illig *et al.*, 2010; Suhre *et al.*, 2011).

***IL3*:** A set of six acylcarnitines, one dicarboxyacylcarnitine, one acyl-alkyl-phosphatidylcholine and one hydroxysphingomyelin was enriched for interleukin 3 (*IL3*) on Biocrates metabolites. *IL3* is known to be a hematopoietic growth factor that stimulates survival, multiplication and differentiation of hematopoietic cells (Lopez *et al.*, 1988). Other studies found that *IL3* stimulates the phospholipid synthesis (Bauer *et al.*, 2005) and suppresses lipid degradation and β-oxidation of fatty acids (Deberardinis *et al.*, 2006). This effect supports the function of *IL3* as hematopoietic growth factor as lipid synthesis is required for proliferation of various cell types (Kuhajda *et al.*, 1994; Deberardinis *et al.*, 2006). In β-oxidation of lipids, e.g. phosphatidylcholines, acylcarnitines are needed for the transport of lipids into the mitochondria. Therefore, the elements of the enriched phenotype set represent the effect of *IL3* on the lipid synthesis. Additionally, a recent study showed that there might be a protective effect of a SNP in *IL3* against malaria attacks (Meyer *et al.*, 2011). GWAS found this gene to be genome wide significantly associated with Chron's disease (Franke *et al.*, 2010).

cytochrome P450 family 4: Four genes of the cytochrome P450 family 4 (*CYP4B1, CYP4A11, CYP4X1, CYP4Z2P*) and one additional gene *KIAA0494* that is overlapping with *CYP4B1* were found on Metabolon metabolites with an enrichment of four slightly different new metabolite sets. The sets included three to six metabolites. Two glycerolipids were part of several sets as well as one fatty acid and two carnitines. The amino acid L-tyrosine and the peptide γ-glutamyltyrosine were part of two enriched new phenotype sets. Also the cofactor heme was identified for two genes. The cytochrome P450 monooxygenase system (CYP) is a multigene superfamily of enzymes that are involved in various reactions e.g. drug metabolism and synthesis of lipids. Heme is a cofactor in these processes (Chaudhary *et al.*, 2009). The metabolites identified as elements of the

metabolite sets seem to reflect the gene product's function. The glycerolipids and the fatty acid might be substrates of the gene products. As stated, heme is a cofactor of the cytochrome P450 enzymes. Carnitines are known to play a role in fatty acid metabolism. Additionally, the gene *CYP4A11* was found to be associated with hypertension (Gainer *et al.*, 2005).

***LINGO2*:** A large set of 17 Metabolon metabolites was identified and significantly enriched at the gene *LINGO2*. The set included various types of metabolites of the lipid metabolism (fatty acids, carnitines, lysolipid, monoacylglycerol), some amino acids, one nucleotide, one peptide (Fibrinogen cleavage peptide) and phenylsulfate. The gene function of leucine rich repeat and Ig domain containing 2 (*LINGO2*) is not known yet. A GWAS identified a genome wide significant association of *LINOG2* with BMI (Speliotes *et al.*, 2010). Moreover, it was reported to be associated with Parkinson's disease (Wu *et al.*, 2011). The elements of the enriched metabolite set possibly hint to an involvement in the metabolism of fatty acids. This could explain the effect on BMI.

These results demonstrated how the identification of new phenotype sets in PSEA improve the knowledge on genes and their function. For *IL3* the elements of the phenotype set are substrates of one gene product's function. Similar observations could be made for the cytochrome P450 family. The elements of the enriched phenotype sets seem to reflect substrates of the processes in which the gene products are involved. As special aspect also heme, which is a known cofactor for these reactions, was identified in the new phenotype set. For these two examples, the findings of PSEA on metabolites supported the known gene functions. In the third example, the enriched phenotype set provided an possibly informative basis for the gene function of *LINGO2*, which is not known yet.

The internal identification of new phenotype sets of PSEA enlarged the approach of testing predefined phenotype sets with valuable hypotheses. One example is the gene *ACADS*. It is known to be strongly associated with medium chain length acyl-carnitines. None of the GGM-defined phenotype set consisted of medium chain length acyl-carnitines and therefore it could not be found with the predefined metabolite sets. But the results of the internal identification of new phenotype sets in PSEA showed an enrichment of a new

phenotype set including medium chain acylcarnitines. The enrichment could be observed for both metabolite panels. This example demonstrates how the internal identification of phenotype sets can help to overcome the limitation given by the selection of predefined phenotype sets.

Most loci but not all that were found with an enrichment of a GGM-defined set also showed an enrichment with a new set. In other words the internal identification of new phenotype sets did not reproduce all enrichments that were found with the predefined phenotype sets. One reason for this is that the GGM-defined phenotype sets could test phenotypes in a more specific way as the internal identification of new phenotype sets. It was observed that the new phenotype sets were larger and included more often metabolites of different metabolite classes as in the GGM-defined metabolite sets. This diversity could be caused by spurious association of metabolites due to the fixed p-value criterion in the identification of phenotype sets. Sets including many spurious associated metabolites would very likely not pass the replication. Due to this some genes might not be identified with an enrichment of a new phenotype set. Therefore, the internal identification of new phenotype sets could not replace the analysis of predefined loci but expand it.

The analysis of new phenotype sets found additional genetic loci that have effect on metabolite sets. Moreover, it was demonstrated that it improved the knowledge on the underlying relations of genes and metabolites and helped to get a more detailed picture of the gene function. Therefore, the extension of PSEA enlarged the possibilities of PSEA. Both testing predefined phenotype sets and identification of new phenotype sets, have special advantages, a combination of both would lead to best results.

6. General Discussion and Conclusion

In the discussion parts of the previous chapters 2-5 specific aspects of PSEA and its applications were discussed. This chapter outlines a more general consideration of the algorithm. Apart from the evaluation of general abilities and limitations, possible analysis strategies including PSEA are presented. At last an overall conclusion is given.

Abilities and Limitations of PSEA

In the introduction two aims of gene based multiple phenotype analysis were named: First the understanding of shared genetic basis of different phenotypes and second the identification of new associated genes. The presented applications of PSEA have demonstrated that PSEA satisfies both objectives. Each significant enrichment of a phenotype set carries information about the common genetic basis of the phenotypes in the set. For example the enrichment of different sets of amino acids in Biocrates and Metabolon data at *CPS1*, as reported in chapter 4, enabled the conclusion of *CPS1* being involved in a shared pathway of these amino acids. Findings in terms of the second aim could be observed for the application to iron and blood count traits in KORA described in chapter 3. PSEA identified new loci associated with various phenotype sets. Some of them were new as they were not genome wide significant in a single phenotype analysis on the same data. This demonstrated the ability of PSEA to identify more loci that have an effect on multiple phenotypes than single phenotype GWAS. The phenotype sets that were enriched at these loci consisted of correlated phenotypes. PSEA exploited this correlation structure, which led to the successful identification of additional loci. The applied permutation test prevented overestimation caused by dependency of phenotypes.

The usage of phenotype sets has several benefits for the analysis of multiple phenotypes. Phenotype sets can be defined easily. Prior knowledge on phenotypes can be integrated, but it is not necessary to have detailed information about their interdependency. Alternative hypotheses can be integrated with different phenotype sets. Phenotype sets might overlap or be a subset of another phenotype set. A single PSEA run can test several different sets in parallel. Another advantage is the ability to deal with many phenotypes and large phenotype sets, which was demonstrated with the successful application to metabolomics data presented in chapter 4 and 5.
Self-evident, the findings of PSEA are highly dependent of the analyzed predefined phenotype sets. In the applications described in chapter 3 and 4, two different approaches were used to determine predefined phenotype sets: the use of prior knowledge in the application to iron and blood count related phenotypes and a data driven strategy with the GGM for the metabolite data sets. The focus of PSEA is to test given sets for enrichment. Therefore, PSEA does not test all possible multiple phenotype combinations. Such a strategy would make the algorithm slower and not applicable to large phenotype sets. Dependent on the situation one can integrate all possible phenotype combinations as predefined sets and force PSEA to test these. But in most situations only selected phenotype sets would be of interest.

The strategy of analyzing predefined phenotype sets is hypotheses testing. The described extension of PSEA that allows the identification of new phenotype sets opens PSEA to be a hypotheses generating application.
For both applications additional significant enrichments were found with this extension. Especially for the application to the metabolites presented in chapter 5 it could be seen that the new identified phenotype sets enabled the identification of new genetic loci. One issue of the fixed p-value criterion for identification of phenotype sets is, that it does not account for the number of phenotypes under consideration. Therefore, more phenotypes lead to more new phenotype sets, in other words a higher false positive rate among the phenotype sets. In the applications one could observe that for the 14 iron and blood phenotypes much fewer sets (296) were identified as for the metabolite panels ($>$ 10,000

each). Apart from the number of phenotypes the strength of association, which was higher for metabolites, was a second reason for the higher number of new phenotype sets on metabolite data. As mentioned above a data driven approach that accounts for number and correlation of phenotypes might improve the results and be a point for further development. But as demonstrated, the fixed p-value selection criterion together with a replication step led to valuable findings that would have been missed if only the predefined sets were analyzed.

It has been discussed above, the usage of GWAS summary statistics in PSEA, as enabled by the second extension, had some limitations. It was seen that the results were dependent on the distribution of single phenotype association test statistics and the LD structure. Even if PSEA on individual level genotypes and phenotypes is favourable, this extension enlarges the applicability of PSEA to situations were individual level genotypes and phenotypes are not available.

PSEA and Analysis Strategies

In this thesis two different situation were presented, for which PSEA was successfully applied. In general, two main kinds of applications could be defined: First, using PSEA as a screening tool for analyzing various phenotype set to get an idea of hidden genetic effects on multiple phenotypes. Second, using PSEA for testing specific phenotype sets based on prior knowledge for genetic effects. Also a mixed form of both types is possible. Using PSEA as screening tool is interesting for example if multiple correlated phenotypes are available. For example in case single phenotype GWAS might reveal shared associations that are not genome wide significant, testing effects on multiple phenotypes with PSEA on a genome wide scale can lead to additional results. In such situations phenotype sets might be defined by correlation or p-value criteria. If one has access to single phenotype GWAS or meta-analysis results but not to individual level genotypes, the extension of PSEA on GWAS results enable the screening for multiple effects without additional analyses on the individual level data. In other cases biological insight can give hypotheses for testing phenotype sets for enrichment in a gene region or on genome wide scale. Often

the identification of a shared genetic basis is the first interest in such situations, but it might also lead to identification of new loci.

To integrate PSEA in an analysis pipeline, on has to distinguish between preparatory analyses and (optional) subsequent analyses based on PSEA results. As stated above, the most important preparation for PSEA is the definition of phenotype sets. In addition to the mentioned approaches, network reconstruction methods or phenotype clustering could be helpful for the definition of phenotype sets. Furthermore, the usage of information on biological relations, e.g. from previous experiments or databases, might be a good source of prior knowledge.
PSEA analyzes if there is any connection between one gene and a set of multiple phenotypes. Subsequent analyses can use the results of PSEA for further investigation. In other words, the screening that is provided by PSEA points to the genetic associations and phenotype sets that are worth an in-deep analyses. Methods that investigate the relation of phenotype sets can be applied subsequently.

Further Development of PSEA

The algorithm of PSEA could be extended to other types of phenotypic input. For example, the usage of binary phenotypes could be integrated with the usage of an logistic regression instead of the linear regression. With the appropriate modifications of the association test PSEA could also be transferred to analysis of related individuals. Moreover, changes in the gene based strategy are thinkable. As mentioned above, a slight modification would change the gene based approach of PSEA to a SNP based algorithm. Even more easily, other definition of genes or genetic loci could be used for PSEA.

PSEA and Pleiotropy

Up to now the term pleiotropy was consciously not mentioned in context with PSEA. Pleiotropy is usually defined as one gene having effects on different independent traits (Sivakumaran *et al.*, 2011). On the contrary, the phenotype sets analyzed by PSEA consist not necessarily of independent phenotypes. PSEA uses the correlation structure of phenotypes in the phenotype sets to identify additionally loci. This was demonstrated

in the analysis of highly correlated sets of iron phenotypes described in the third chapter. The strategy of the permutation test prevents overestimation of effects on correlated phenotypes, as the correlation structure is conserved in the permutated phenotypes. If the phenotype sets are made up of independent phenotypes, PSEA will investigate pleiotropic effects on these phenotypes. Therefore, on the one hand PSEA can be used for the analysis of pleiotropic effects, but on the other hand is not limited to this.

Conclusion

Genetic tests become more and more popular in the health care system and in direct-to-customer services. Therefore, it is increasingly important to unravel shared genetic effects on different traits. Hidden association may contain important information on the relation and development of disease. PSEA was demonstrated to improve our understanding of the genetic connection of phenotypes and to identify new loci. Moreover, it enables both testing of hypotheses by the usage of prior knowledge in predefined phenotype sets and generation of new hypotheses with identification of new sets. It is easy to apply and can be generalized for other questions. Therefore, PSEA is a valuable tool for the understanding of shared genetic basis of phenotypes and diseases. It can help to improve the interpretation of genetic tests, our understanding of diseases and possibilities of therapy development.

A. Supplementary Information: Methods

A.1. Permutation Strategy of PSEA

For a setting with N individuals and a phenotype set of M phenotypes let P be the matrix of all phenotypes in the set.

$$P = \begin{pmatrix} p_{1,1} & \cdots & p_{1,M} \\ p_{2,1} & \cdots & p_{2,M} \\ \vdots & \ddots & \vdots \\ p_{N,1} & \cdots & p_{N,M} \end{pmatrix} \tag{A.1}$$

Let $i \in \{1, \ldots, N\}$ and $m \in \{1, \ldots, M\}$. $p_{i,m}$ is the value of the phenotype m for the person i. The vector $p_{i,\cdot} = (p_{i,1}, p_{i,2}, \ldots, p_{i,M})$ denotes the values of all phenotypes in the phenotype set for the person i.

In the permutation these phenotype vectors are permuted over all individuals. With this the phenotype values of a person stay together in the permutations, only the asignment to the individuals is modified. In other words the permutation is performed by interchanging the rows of the matrix P. For a permutation $j = (j_1, j_2, ...j_N)$ of the vector $(1, 2, \ldots, N)$ the matrix of permuted phenotypes is:

$$P^{(j)} = \begin{pmatrix} p_{j_1,1} & \cdots & p_{j_1,M} \\ p_{j_2,1} & \cdots & p_{j_2,M} \\ \vdots & \ddots & \vdots \\ p_{j_N,1} & \cdots & p_{j_N,M} \end{pmatrix} \tag{A.2}$$

In mathematical terms this permutation is equivalent to multiplication of the permutation matrix $\Pi^{(j)} = \{e_{j_1}, \ldots, e_{j_N}\}$ with e_{j_k}, the j_kth unit vektor and $k \in \{1, \ldots, N\}$, to the left side of the phenotype matrix P:

$$P^{(j)} = \Pi^{(j)} \cdot P. \tag{A.3}$$

This permutation scheme uses only the phenotypes of the phenotype set. Therefore, it can be called *self-contained* in contrast to approaches that include also other phenotypes than those of the phenotype set under consideration.

A.2. Estimation with FDR and FWER

The ESs for different phenotype sets were not comparable as the phenotype sets differed in number of phenotypes and correlation between the phenotypes. Moreover, larger genes were more likely to have a high gene based statistic by chance than smaller genes, since larger genes include a higher number of SNPs. The ES can be made comparable by standardization with mean and standard deviation of all permutation based ES. The NES was calculated as

$$NES(PS, gene) = \frac{ES\left(PS, gene\right) - mean_j\left(ES^{(j)}\left(PS, gene\right)\right)}{sd_j\left(ES^{(j)}\left(PS, gene\right)\right)} \ . \tag{A.4}$$

This standardization method was used in various gene set enrichment approaches (Guo *et al.*, 2009; Wang *et al.*, 2007). Replacing ES in the numerator of equation (A.4) by a permutation based ES ($ES^{(j)}(PS, gene)$) resulted in the permutation based NES ($NES^{(j)}(PS, gene)$).

With this the FDR and FWER can be used for estimation of enrichment.

False discovery rate (FDR): The FDR controls the fraction of false positive findings (Benjamini and Hochberg, 1995). According to the FDR used by Wang *et al.* (2007), the FDR for a phenotype set (PS*) and a gene (gene*) taking into account all other

phenotype sets and genes was

$$FDR\left(PS^*, gene^*\right) = \frac{\#_{k,j,i}\left[NES\left(PS^*, gene^*\right) \leq NES^{(j)}\left(PS_k, gene_i\right)\right]}{\#_{k,i}\left[NES\left(PS^*, gene^*\right) \leq NES\left(PS_k, gene_i\right)\right] \times N_{perm}} \ . \quad \text{(A.5)}$$

Thereby, the index j indicates the permutation ($1 \leq j \leq N_{perm}$), index i the gene ($1 \leq i \leq N_{gene}$) and index k the phenotype set ($1 \leq k \leq N_{pset}$).

Family wise error rate (FWER): The FWER is a highly conservative correction procedure (Benjamini and Hochberg, 1995). It refers to the probability that the results contain one or more false positive results. We transferred the FWER application of Wang *et al.* (2007) in gene set enrichment to the situation of PSEA:

$$FWER\left(PS^*, gene^*\right) = \frac{\#_j\left[NES\left(PS^*, gene^*\right) \leq \max_{k,i}\left(NES^{(j)}\left(PS_k, gene_i\right)\right)\right]}{N_{perm}} \ . \quad \text{(A.6)}$$

Thereby, the index j indicates the permutation ($1 \leq j \leq N_{perm}$), index i the gene ($1 \leq i \leq N_{gene}$) and index k the phenotype set ($1 \leq k \leq N_{pset}$).

A.3. Level and Power of the PSEA test

Level

Several publications discuss the problem of inflated α error rate in permutation tests (Huang *et al.*, 2006; Kaizar *et al.*, 2011). They describe permutation tests in the situation of a mean comparison with different distributions of characteristics in the groups, which can lead to an inflation of the α error rate. In the presented application of PSEA the effect of a SNP is assumed to be additive, which is the standard model in genome wide association analyses. With this, the genotype could be treated as continuous variable. Therefore, we think that this model avoids problems as mentioned above. A more detailed discussion of suchlike issues might be needed if PSEA is generalized to other genetic models.

Another point that should be mentioned in this context is the effect of a limited number of permutations. As discussed in chapter 2 due to the limited number of permutations it

is not always possible to apply Bonferroni correction and this could lead to an increased risk of false positive findings. As it is not always feasible to increase the number of permutations, it was stated that a replication in an independent cohort should be performed. The presented applications demonstrated that PSEA identified enrichments of phenotype sets at genes that could be replicated.

Power

The theoretical determination of the power of PSEA is not feasible, as the testing procedure is very complex. A simulation based determination of power would also mean extensive analyses as there are many variable parameters that affect the performance. The most important parameters are: The mapping of SNPs to genes (number, LD structure), as well as the association of SNPs and the number of phenotypes and their correlation structure. General conclusions that consider all aspects could therefore only be gained with extensive simulations.

Instead of a theoretical discussion of power it was demonstrated in the presented applications (chapter 3, 4 and 5) that the PSEA algorithm is able to detect interesting associations that could be replicated in independent cohorts. Moreover, several phenotype sets were found to be enriched at genes that are biologically related to the phenotypes. Those findings were not only replicated, but biological plausible.

In chapter 3 the findings of PSEA were compared with results of single phenotype GWAS, which is the state-of-the-art procedure for analysing associations of SNPs and (single) phenotypes. It was exemplarily demonstrated that some genes could be identified with PSEA for an enrichment of a phenotype set, but all single phenotype associations of the set elements at this gene had a p-value above the genome wide significance level. As the genes were also found and replicated in independent large meta-analyses, they are likely no false positives. Therefore, it was concluded that the power of PSEA for these genes was higher than the single GWAS approach. Of course this is only one example and therefore of limited information, but nevertheless it demonstrated the ability of PSEA to detect other genes than single phenotype GWAS do.

B. Supplementary Information: Blood Iron Phenotypes

Table B.1.: **Information about blood and iron phenotypes measurements** in KORA F3 and KORA F4.

phenotype	unit	method	
		KORA F3	KORA F4
Iron	$\mu mol/l$	Colorimetric assay, (Cobas cr, Roche)	
Ferritin	ng/ml	Electrochemiluminescence immunoassay (ECLIA)	
stfr	mg/l	Tina-quant cr immunoturbidometry (Roche)	
Transferrin	g/l	Immunonephelometry, (Behring Nephelometer cr, Siemens)	
Ferritin	ng/ml	electrochemiluminescence immunoassay (Roche)	
Transferrin saturation	$\mu mol/g$	3.98 * Serum-Fe [$\mu mol/l$] / Transferrin [g/l]	
HCT	l/l	Beckman Coulter STKS	Beckman Coulter LH750
HGB	g/l	Beckman Coulter STKS	Beckman Coulter LH750
MCH	pg	Beckman Coulter STKS	Beckman Coulter LH750
MCHC	g/l	Beckman Coulter STKS	Beckman Coulter LH750
MCV	fl	Beckman Coulter STKS	Beckman Coulter LH750
RBC	$/pl$	Beckman Coulter STKS	Beckman Coulter LH750
MPV	fl	Beckman Coulter STKS	Beckman Coulter LH750
PLT	$/nl$	Beckman Coulter STKS	Beckman Coulter LH750
WBC	$/nl$	Beckman Coulter STKS	Beckman Coulter LH750

stfr: soluble transferrin receptor

Table B.2.: **Population statistics of blood and iron phenotypes.** All characteristics were calculated after exclusion of extreme phenotype values (±3 standard deviations from mean). Units of phenotypes as specified in Table B.1. (N: absolute number of individuals, min: minimal measurement, max: maximal measurement, sd: standard deviation).

phenotype	KORA F3					KORA F4				
	N	min	max	mean	sd	N	min	max	mean	sd
Iron	1625	2.3	32.8	16.67	5.12	1790	3.6	41.4	20.9	6.27
Ferritin	1610	3.9	829.5	199.04	162.56	1770	2.51	849.8	184.51	147.78
stfr	1620	1.12	5.57	2.67	0.72	1790	1.15	6.86	2.95	0.75
Transferrin	1622	1.55	3.56	2.50	0.33	1794	1.54	3.597	2.53	0.34
Transferrin saturation	1624	2.87	54.70	26.81	8.73	1788	4.21	69.32	33.38	10.82
HCT	1633	0.32	0.52	0.42	0.03	1803	0.32	0.51	0.41	0.03
HGB	1631	106	176	141.81	11.44	1805	106	176	141.10	11.53
MCH	1620	26.2	35.8	30.89	1.44	1787	26.3	36.1	31.2	1.53
MCHC	1633	317	357	337.06	6.78	1807	319	363	341.04	7.27
MCV	1625	79.4	104.1	91.55	3.79	1792	78.6	104.1	91.44	4
RBC	1635	3.41	5.73	4.6	0.39	1801	3.4	5.67	4.53	0.37
MPV	1600	6.4	11.4	8.67	0.91	1791	6.4	11.7	8.86	0.89
PLT	1625	75	421	243.76	54.04	1796	75	433	250.81	57.22
WBC	1639	3.2	13.4	6.78	1.66	1790	2.7	10.8	5.82	1.46

stfr: soluble transferrin receptor

Table B.3.: **Published genes with known association with iron and/or blood traits.** The presented information was extracted from the corresponding publications ([Benyamin, et al. 2009; Chambers, et al. 2009; Kullo, et al. 2010; Oexle, et al. 2011; Soranzo, et al. 2009; Tanaka, et al. 2010]). In the last column all genes are specified to which the published SNP was be mapped using the definition of gene regions (110 kb upstream, 40 kb downstream).

SNP	trait	gene	p-value	N	publication	other genes that include the published SNP
rs10914144	MPV	DNM3	2.10E-14	13943	Soranzo *et al.* (2009)	DNM3
rs11065987	PLT	ATXN2	2.20E-13	13943	Soranzo *et al.* (2009)	ACAD10, ATXN2, BRAP
rs11066301	PLT	PTPN11	7.70E-12	13943	Soranzo *et al.* (2009)	RPL6, PTPN11, C12orf51
rs11071720	MPV	TPM1	1.90E-08	13943	Soranzo *et al.* (2009)	TPM1, LACTB
rs11602954	MPV	BET1L	1.30E-14	13943	Soranzo *et al.* (2009)	ATHL1, ODF3, IFITM2, LOC100133161, BET1L, SCGB1C1, SIRT3, RIC8A, PSMD13, NLRP6
rs11970772	MCV	BYSL, CCND3	7.00E-19	13943	Soranzo *et al.* (2009)	TAF8, CCND3, BYSL, USP49, MED20

rs12485738	MPV	ARHGEF3	5.50E-31	13943	Soranzo *et al.* (2009)	ARHGEF3
rs1668873	MPV	TMCC2	1.40E-20	13943	Soranzo *et al.* (2009)	DSTYK, NUAK2, TMCC2
rs17342717	MCH	SLC17A1	4.66E-08	3012	Kullo *et al.* (2010)	SLC17A3, HIST1H2AA, SLC17A1
rs17609240	WBC	GSDMA, ORMDL3	4.66E-08	13943	Soranzo *et al.* (2009)	IKZF3, GSDMA, SNORD124, THRA, ORMDL3, GSDMB, PSMD3, CSF3
rs1800562	MCV	HFE	1.40E-23	13943	Soranzo *et al.* (2009)	HIST1H2BF, HIST1H1T, HIST1H2BC, HIST1H2BB, HIST1H2BE, HIST1H2BD, HIST1H1C, HFE, HIST1H1E, HIST1H2AC, HIST1H1A, HIST1H2AB, HIST1H4C, HIST1H4B, HIST1H3B
	stfr		1.17E-09	6616	Oexle *et al.* (2011)	
	MCH		2.76E-09	3012	Kullo *et al.* (2010)	
	transferrin		1.10E-10	459	Benyamin *et al.* (2009)	
	trans. sat.		1.50E-15	459	Benyamin *et al.* (2009)	
	iron		3.50E-11	459	Benyamin *et al.* (2009)	
rs210135	PLT	BAK1	3.70E-10	13943	Soranzo *et al.* (2009)	ITPR3, GGNBP1, BAK1, C6orf227
rs2138852	MPV	TAOK1	1.40E-23	13943	Soranzo *et al.* (2009)	NUFIP2, TAOK1
rs236918	stfr	PCSK7	1.41E-27	6616	Oexle *et al.* (2011)	SIDT2, CEP164, RNF214, PCSK7, TAGLN
rs2393967	MPV	JMJD1C	3.30E-21	13943	Soranzo *et al.* (2009)	MIR1296, LOC84989, JMJD1C
rs342293	MPV	PIK3CG	1.60E-33	13943	Soranzo *et al.* (2009)	FLJ36031
rs3811647	transferrin	TF	3.00E-15	459	Benyamin *et al.* (2009)	SRPRB, TOPBP1, TF
rs385893	PLT	AK3	8.50E-17	13943	Soranzo *et al.* (2009)	AK3, MIR101-2, RCL1, C9orf68
rs4895441	RBC	HBS1L	3.12E-14	3012	Kullo *et al.* (2010)	HBS1L, MYB
	MCH	MYB	5.03E-13	3012	Kullo *et al.* (2010)	
	MCV		2.46E-13	3012	Kullo *et al.* (2010)	
rs6136489	MPV	SIRPA	7.70E-11	13943	Soranzo *et al.* (2009)	SIRPA, PDYN
rs647316	MPV	EHD3	3.20E-11	13943	Soranzo *et al.* (2009)	CAPN14, GALNT14, EHD3
rs7385804	RBC	TFR2	4.90E-10	13943	Soranzo *et al.* (2009)	PCOLCE, FBXO24, POP7, GNB2, ZAN, SAP25, MOSPD3, ACTL6B, LRCH4, EPO, TFR2
rs7961894	mpv	WDR66	2.70E-44	13943	Soranzo *et al.* (2009)	HPD, PSMD9, WDR66, BCL7A
rs855791	mchc	TMPRSS6	1.10E-12	3012	Kullo *et al.* (2010)	C22orf33, TST, TMPRSS6, KCTD17, MPST
	stfr		1.69E-15	6616	Oexle *et al.* (2011)	
	mcv		5.41E-09	3012	Kullo *et al.* (2010)	
	hgb		1.60E-13	16001	Chambers *et al.* (2009)	
rs4820268	iron	TMPRSS6	5.12E-09	2488	Kullo *et al.* (2010)	C22orf33, TST, KCTD17, TMPRSS6
	mch		2.41E-11	3012	Kullo *et al.* (2010)	
rs893001	mpv	CD226	1.40E-10	13943	Soranzo *et al.* (2009)	CD226, DOK6
rs9609565	mcv	FBXO7	4.30E-10	13943	Soranzo *et al.* (2009)	BPIL2, C22orf28, FBXO7, RFPL3S

stfr: soluble transferrin receptor, trans. sat.: transferrin saturation

Table B.4.: **Number of genes that were identified by PSEA on pruned and not pruned SNPs.** The table compares the number of genes identified with PSEA based on pruned and not pruned SNPs (KORA F4). The counts for PSEA on individual genotypes and phenotype measurements (PSEA-GENO) and for PSEA based on GWAS results (PSEA-GWAS) are presented. For better comparability genes that were not represented in the pruned data are also excluded form the counts for not-pruned data.

	PSEA-GENO				PSEA-GWAS			
	KORA F4 not pruned number of genes		KORA F4 pruned number of genes		KORA F4 not pruned number of genes		KORA F4 pruned number of genes	
	identified	previous	identified	previous	identified	previous	identified	previous
gene	269	43	274	43	56	22	159	34
gene group	70	15	71	15	16	7	48	12

Table B.5.: **Number of genes that were identified by PSEA on pruned SNPs.** This table compares the number of identified genes per predefined phenotype set and the number of previously published genes among those for PSEA on genotypes and phenotypes (PSEA-GENO) and PSEA on GWAS results (PSEA-GWAS) for the pruned SNP data.

	PSEA on pruned data					
	number of identified genes			number of previously associated genes		
	PSEA-GENO	PSEA-GWAS	percentage	PSEA-GENO	PSEA-GWAS	percentage
set_blood1	86	22	26 %	9	5	56 %
set_blood2	66	36	55 %	9	6	67 %
set_blood3	22	14	64 %	10	8	80 %
set_iron1	81	39	48 %	22	8	36 %
set_iron2	87	26	30 %	25	7	28 %
set_iron3	73	37	51 %	15	15	100 %

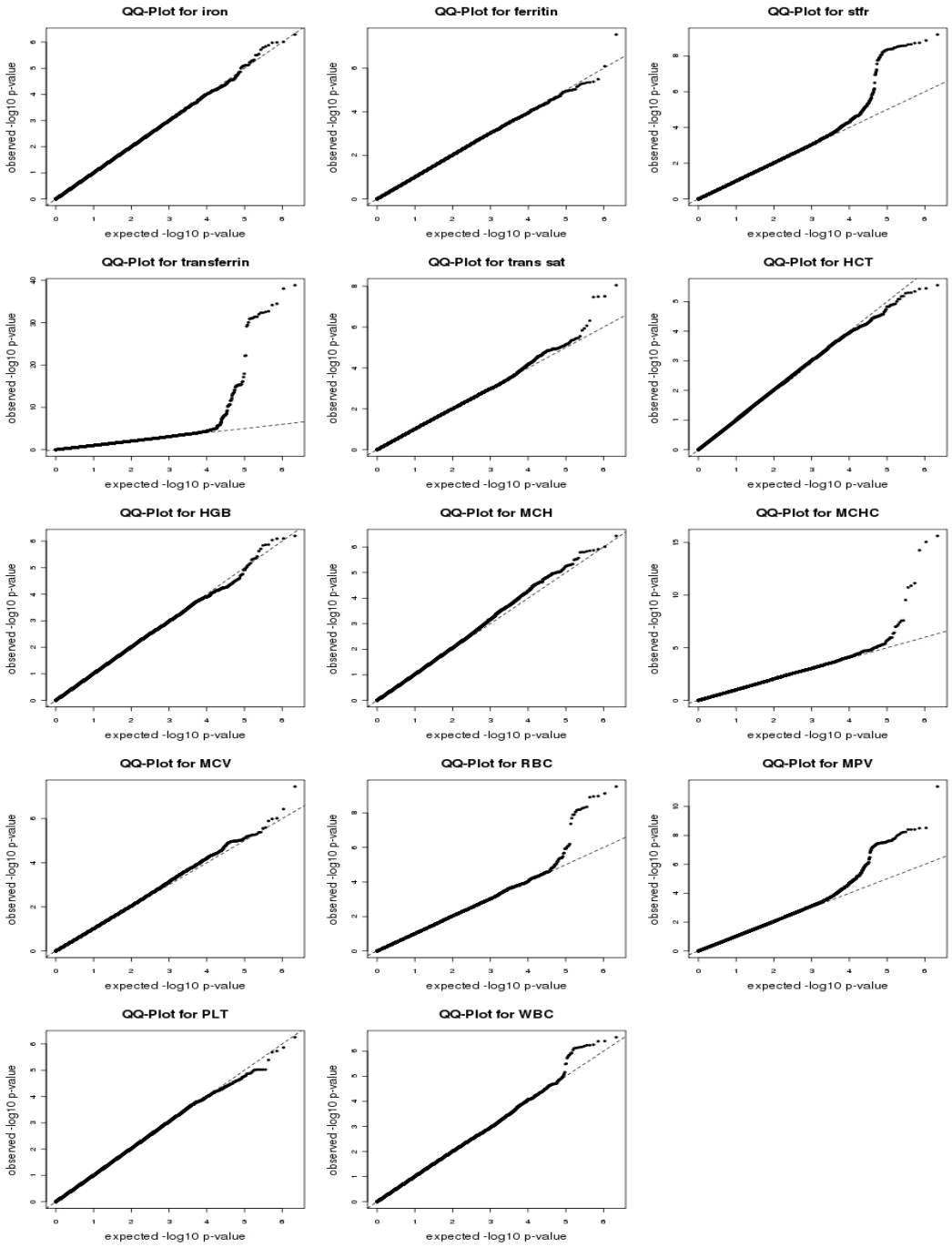


Figure B.1.: **QQ-plots of KORA F4 iron and blood trait GWAS results.** GWAS calculated with SNPTEST on the residuals concerning sex and age of all analyzed iron and blood phenotypes in KORA F4 (stfr: soluble transferrin receptor, trans sat: transferrin saturation).

C. Supplementary Information: Metabolomics Data

Table C.1.: **Information on Biocrates metabolites.** For each metabolite included in the analyses the short name, the full biological name and information about the metabolites class is given.

short name	full biochemical name	pathway
H1	hexose	sugar
Arg	arginine	amino acids
Gln	glutamine	amino acids
Gly	glycine	amino acids
His	histidine	amino acids
Met	methionine	amino acids
Orn	ornithine	amino acids
Phe	phenylalanine	amino acids
Pro	proline	amino acids
Ser	serine	amino acids
Thr	threonine	amino acids
Trp	tryptophan	amino acids
Tyr	tyrosine	amino acids
Val	valine	amino acids
xLeu	leucine/isoleucine	amino acids
C0	carnitine	carnitine
C2	acetylcarnitine	acylcarnitines
C3	propionylcarnitine	acylcarnitines
C4	butyrylcarnitine	acylcarnitines
C5	valerylcarnitine	acylcarnitines
C5.1	tiglylcarnitine	acylcarnitines
C6.1	hexenoylcarnitine	acylcarnitines
C8	octanoylcarnitine	acylcarnitines
C8.1	octenoylcarnitine	acylcarnitines
C9	nonaylcarnitine	acylcarnitines
C10	decanoylcarnitine	acylcarnitines
C10.1	decenoylcarnitine	acylcarnitines
C10.2	decadienylcarnitine	acylcarnitines
C12	dodecanoylcarnitine	acylcarnitines
C12.1	dodecenoylcarnitine	acylcarnitines
C14	tetradecanoylcarnitine	acylcarnitines
C14.1	tetradecenoylcarnitine	acylcarnitines

C14.2	tetradecadienylcarnitine	acylcarnitines
C16	hexadecanoylcarnitine	acylcarnitines
C16.1	hexadecenoylcarnitine	acylcarnitines
C18	octadecanoylcarnitine	acylcarnitines
C18.1	octadecenoylcarnitine	acylcarnitines
C18.2	octadecadienylcarnitine	acylcarnitines
C3.DC...C4.OH	hydroxybutyrylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C5.DC...C6.OH	glutarylcarnitine (hydroxyhexanoylcarnitine)	hydroxy- and dicarboxy-acylcarnitines
C5.M.DC	methylglutarylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C3.DC.M...C5.OH	hydroxyvalerylcarnitine (methylmalonylcarnitine)	hydroxy- and dicarboxy-acylcarnitines
C5.1.DC	glutaconylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C4.1.DC...C6	hexanoylcarnitine (fumarylcarnitine)	hydroxy- and dicarboxy-acylcarnitines
C7.DC	pimeloylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C12.DC	dodecanedioylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C14.1.OH	hydroxytetradecenoylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C14.2.OH	hydroxytetradecadienylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C16.1.OH	hydroxyhexadecenoylcarnitine	hydroxy- and dicarboxy-acylcarnitines
C16.2.OH	hydroxyhexadecadienylcarnitine	hydroxy- and dicarboxy-acylcarnitines
SM.C16.0	sphingomyeline C16:0	sphingomyelins
SM.C16.1	sphingomyeline C16:1	sphingomyelins
SM.C18.0	sphingomyeline C18:0	sphingomyelins
SM.C18.1	sphingomyeline C18:1	sphingomyelins
SM.C20.2	sphingomyeline C20:2	sphingomyelins
SM.C24.0	sphingomyeline C24:0	sphingomyelins
SM.C24.1	sphingomyeline C24:1	sphingomyelins
SM.C26.0	sphingomyeline C26:0	sphingomyelins
SM.C26.1	sphingomyeline C26:1	sphingomyelins
SM..OH..C14.1	hydroxysphingomyeline C14:1	hydroxysphingomyelins
SM..OH..C16.1	hydroxysphingomyeline C16:1	hydroxysphingomyelins
SM..OH..C22.1	hydroxysphingomyeline C22:1	hydroxysphingomyelins
SM..OH..C22.2	hydroxysphingomyeline C22:2	hydroxysphingomyelins
SM..OH..C24.1	hydroxysphingomyeline C24:1	hydroxysphingomyelins
PC.aa.C24.0	phosphatidylcholine diacyl C24:0	diacyl-phosphatidylcholines
PC.aa.C26.0	phosphatidylcholine diacyl C26:0	diacyl-phosphatidylcholines
PC.aa.C28.1	phosphatidylcholine diacyl C28:1	diacyl-phosphatidylcholines
PC.aa.C30.0	phosphatidylcholine diacyl C30:0	diacyl-phosphatidylcholines
PC.aa.C32.0	phosphatidylcholine diacyl C32:0	diacyl-phosphatidylcholines
PC.aa.C32.1	phosphatidylcholine diacyl C32:1	diacyl-phosphatidylcholines
PC.aa.C32.2	phosphatidylcholine diacyl C32:2	diacyl-phosphatidylcholines
PC.aa.C32.3	phosphatidylcholine diacyl C32:3	diacyl-phosphatidylcholines
PC.aa.C34.1	phosphatidylcholine diacyl C34:1	diacyl-phosphatidylcholines
PC.aa.C34.2	phosphatidylcholine diacyl C34:2	diacyl-phosphatidylcholines
PC.aa.C34.3	phosphatidylcholine diacyl C34:3	diacyl-phosphatidylcholines
PC.aa.C34.4	phosphatidylcholine diacyl C34:4	diacyl-phosphatidylcholines
PC.aa.C36.0	phosphatidylcholine diacyl C36:0	diacyl-phosphatidylcholines
PC.aa.C36.1	phosphatidylcholine diacyl C36:1	diacyl-phosphatidylcholines
PC.aa.C36.2	phosphatidylcholine diacyl C36:2	diacyl-phosphatidylcholines
PC.aa.C36.3	phosphatidylcholine diacyl C36:3	diacyl-phosphatidylcholines
PC.aa.C36.4	phosphatidylcholine diacyl C36:4	diacyl-phosphatidylcholines
PC.aa.C36.5	phosphatidylcholine diacyl C36:5	diacyl-phosphatidylcholines
PC.aa.C36.6	phosphatidylcholine diacyl C36:6	diacyl-phosphatidylcholines
PC.aa.C38.0	phosphatidylcholine diacyl C38:0	diacyl-phosphatidylcholines
PC.aa.C38.3	phosphatidylcholine diacyl C38:3	diacyl-phosphatidylcholines
PC.aa.C38.4	phosphatidylcholine diacyl C38:4	diacyl-phosphatidylcholines
PC.aa.C38.5	phosphatidylcholine diacyl C38:5	diacyl-phosphatidylcholines
PC.aa.C38.6	phosphatidylcholine diacyl C38:6	diacyl-phosphatidylcholines
PC.aa.C40.1	phosphatidylcholine diacyl C40:1	diacyl-phosphatidylcholines
PC.aa.C40.2	phosphatidylcholine diacyl C40:2	diacyl-phosphatidylcholines
PC.aa.C40.3	phosphatidylcholine diacyl C40:3	diacyl-phosphatidylcholines

PC.aa.C40.4	phosphatidylcholine diacyl C40:4	diacyl-phosphatidylcholines
PC.aa.C40.5	phosphatidylcholine diacyl C40:5	diacyl-phosphatidylcholines
PC.aa.C40.6	phosphatidylcholine diacyl C40:6	diacyl-phosphatidylcholines
PC.aa.C42.0	phosphatidylcholine diacyl C42:0	diacyl-phosphatidylcholines
PC.aa.C42.1	phosphatidylcholine diacyl C42:1	diacyl-phosphatidylcholines
PC.aa.C42.2	phosphatidylcholine diacyl C42:2	diacyl-phosphatidylcholines
PC.aa.C42.4	phosphatidylcholine diacyl C42:4	diacyl-phosphatidylcholines
PC.aa.C42.5	phosphatidylcholine diacyl C42:5	diacyl-phosphatidylcholines
PC.aa.C42.6	phosphatidylcholine diacyl C42:6	diacyl-phosphatidylcholines
PC.ae.C30.0	phosphatidylcholine acyl-alkyl C30:0	acyl-alkyl-phosphatidylcholines
PC.ae.C30.2	phosphatidylcholine acyl-alkyl C30:2	acyl-alkyl-phosphatidylcholines
PC.ae.C32.1	phosphatidylcholine acyl-alkyl C32:1	acyl-alkyl-phosphatidylcholines
PC.ae.C32.2	phosphatidylcholine acyl-alkyl C32:2	acyl-alkyl-phosphatidylcholines
PC.ae.C34.0	phosphatidylcholine acyl-alkyl C34:0	acyl-alkyl-phosphatidylcholines
PC.ae.C34.1	phosphatidylcholine acyl-alkyl C34:1	acyl-alkyl-phosphatidylcholines
PC.ae.C34.2	phosphatidylcholine acyl-alkyl C34:2	acyl-alkyl-phosphatidylcholines
PC.ae.C34.3	phosphatidylcholine acyl-alkyl C34:3	acyl-alkyl-phosphatidylcholines
PC.ae.C36.0	phosphatidylcholine acyl-alkyl C36:0	acyl-alkyl-phosphatidylcholines
PC.ae.C36.1	phosphatidylcholine acyl-alkyl C36:1	acyl-alkyl-phosphatidylcholines
PC.ae.C36.2	phosphatidylcholine acyl-alkyl C36:2	acyl-alkyl-phosphatidylcholines
PC.ae.C36.3	phosphatidylcholine acyl-alkyl C36:3	acyl-alkyl-phosphatidylcholines
PC.ae.C36.4	phosphatidylcholine acyl-alkyl C36:4	acyl-alkyl-phosphatidylcholines
PC.ae.C36.5	phosphatidylcholine acyl-alkyl C36:5	acyl-alkyl-phosphatidylcholines
PC.ae.C38.0	phosphatidylcholine acyl-alkyl C38:0	acyl-alkyl-phosphatidylcholines
PC.ae.C38.1	phosphatidylcholine acyl-alkyl C38:1	acyl-alkyl-phosphatidylcholines
PC.ae.C38.2	phosphatidylcholine acyl-alkyl C38:2	acyl-alkyl-phosphatidylcholines
PC.ae.C38.3	phosphatidylcholine acyl-alkyl C38:3	acyl-alkyl-phosphatidylcholines
PC.ae.C38.4	phosphatidylcholine acyl-alkyl C38:4	acyl-alkyl-phosphatidylcholines
PC.ae.C38.5	phosphatidylcholine acyl-alkyl C38:5	acyl-alkyl-phosphatidylcholines
PC.ae.C38.6	phosphatidylcholine acyl-alkyl C38:6	acyl-alkyl-phosphatidylcholines
PC.ae.C40.0	phosphatidylcholine acyl-alkyl C40:0	acyl-alkyl-phosphatidylcholines
PC.ae.C40.1	phosphatidylcholine acyl-alkyl C40:1	acyl-alkyl-phosphatidylcholines
PC.ae.C40.2	phosphatidylcholine acyl-alkyl C40:2	acyl-alkyl-phosphatidylcholines
PC.ae.C40.3	phosphatidylcholine acyl-alkyl C40:3	acyl-alkyl-phosphatidylcholines
PC.ae.C40.4	phosphatidylcholine acyl-alkyl C40:4	acyl-alkyl-phosphatidylcholines
PC.ae.C40.5	phosphatidylcholine acyl-alkyl C40:5	acyl-alkyl-phosphatidylcholines
PC.ae.C40.6	phosphatidylcholine acyl-alkyl C40:6	acyl-alkyl-phosphatidylcholines
PC.ae.C42.0	phosphatidylcholine acyl-alkyl C42:0	acyl-alkyl-phosphatidylcholines
PC.ae.C42.1	phosphatidylcholine acyl-alkyl C42:1	acyl-alkyl-phosphatidylcholines
PC.ae.C42.2	phosphatidylcholine acyl-alkyl C42:2	acyl-alkyl-phosphatidylcholines
PC.ae.C42.3	phosphatidylcholine acyl-alkyl C42:3	acyl-alkyl-phosphatidylcholines
PC.ae.C42.4	phosphatidylcholine acyl-alkyl C42:4	acyl-alkyl-phosphatidylcholines
PC.ae.C42.5	phosphatidylcholine acyl-alkyl C42:5	acyl-alkyl-phosphatidylcholines
PC.ae.C44.3	phosphatidylcholine acyl-alkyl C44:3	acyl-alkyl-phosphatidylcholines
PC.ae.C44.4	phosphatidylcholine acyl-alkyl C44:4	acyl-alkyl-phosphatidylcholines
PC.ae.C44.5	phosphatidylcholine acyl-alkyl C44:5	acyl-alkyl-phosphatidylcholines
PC.ae.C44.6	phosphatidylcholine acyl-alkyl C44:6	acyl-alkyl-phosphatidylcholines
lysoPC.a.C14.0	lysophosphatidylcholine acyl C14:0	lyso-phosphatidylcholines
lysoPC.a.C16.0	lysophosphatidylcholine acyl C16:0	lyso-phosphatidylcholines
lysoPC.a.C16.1	lysophosphatidylcholine acyl C16:1	lyso-phosphatidylcholines
lysoPC.a.C17.0	lysophosphatidylcholine acyl C17:0	lyso-phosphatidylcholines
lysoPC.a.C18.0	lysophosphatidylcholine acyl C18:0	lyso-phosphatidylcholines
lysoPC.a.C18.1	lysophosphatidylcholine acyl C18:1	lyso-phosphatidylcholines
lysoPC.a.C18.2	lysophosphatidylcholine acyl C18:2	lyso-phosphatidylcholines
lysoPC.a.C20.3	lysophosphatidylcholine acyl C20:3	lyso-phosphatidylcholines
lysoPC.a.C20.4	lysophosphatidylcholine acyl C20:4	lyso-phosphatidylcholines
lysoPC.a.C24.0	lysophosphatidylcholine acyl C24:0	lyso-phosphatidylcholines
lysoPC.a.C26.1	lysophosphatidylcholine acyl C26:1	lyso-phosphatidylcholines
lysoPC.a.C28.0	lysophosphatidylcholine acyl C28:0	lyso-phosphatidylcholines
lysoPC.a.C28.1	lysophosphatidylcholine acyl C28:1	lyso-phosphatidylcholines

Table C.2.: **Information on Metabolon metabolites.** For each metabolite included in the analyses the short name, a long name and the information about super- and subpathway is given.

short name	full biochemical name	superpathway	subpathway
ALANINE	alanine	amino acid	alanine and aspartate metabolism
ALPHAHYD	alpha-hydroxyisovalerate	amino acid	valine, leucine and isoleucine metabolism
ARGININE	arginine	amino acid	urea cycle; arginine-, proline-, metabolism
ASPARAGI	asparagine	amino acid	alanine and aspartate metabolism
ASPARTAT	aspartate	amino acid	alanine and aspartate metabolism
BETAINE	betaine	amino acid	glycine, serine and threonine metabolism
CGLYCOSY	c-glycosyltryptophan	amino acid	tryptophan metabolism
CITRULLI	citrulline	amino acid	urea cycle; arginine-, proline-, metabolism
CREATINE	creatine	amino acid	creatine metabolism
CREATINI	creatinine	amino acid	creatine metabolism
CYSTEINE	cysteine	amino acid	cysteine, methionine, SAM, taurine metabolism
DIMETHYL	dimethylarginine	amino acid	urea cycle; arginine-, proline-, metabolism
GLUTAMAT	glutamate	amino acid	glutamate metabolism
GLUTAMIN	glutamine	amino acid	glutamate metabolism
GLYCINE	glycine	amino acid	glycine, serine and threonine metabolism
HISTIDIN	histidine	amino acid	histidine metabolism
INDOLEAC	indoleacetate	amino acid	tryptophan metabolism
INDOLEPR	indolepropionate	amino acid	tryptophan metabolism
ISOLEUCI	isoleucine	amino acid	valine, leucine and isoleucine metabolism
KYNURENI	kynurenine	amino acid	tryptophan metabolism
LEUCINE	leucine	amino acid	valine, leucine and isoleucine metabolism
LYSINE	lysine	amino acid	lysine metabolism
METHIONI	methionine	amino acid	cysteine, methionine, SAM, taurine metabolism
NACETYLA	N-acetylornithine	amino acid	valine, leucine and isoleucine metabolism
NACETYLO	N-acetylornithine	amino acid	urea cycle; arginine-, proline-, metabolism
ORNITHIN	ornithine	amino acid	urea cycle; arginine-, proline-, metabolism
PCRESOLS	p-cresol	amino acid	phenylalanine and tyrosine metabolism
PHENYLAL	phenylalanine	amino acid	phenylalanine and tyrosine metabolism
PIPECOLA	pipecolate	amino acid	lysine metabolism
PROLINE	proline	amino acid	urea cycle; arginine-, proline-, metabolism
PYROGLUT	pyroglutamine	amino acid	glutamate metabolism
SERINE	serine	amino acid	glycine, serine and threonine metabolism
SEROTONI	serotonin	amino acid	tryptophan metabolism
STACHYDR	stachydrine	amino acid	urea cycle; arginine-, proline-, metabolism
THREONIN	threonine	amino acid	glycine, serine and threonine metabolism
TRANS4HY	hydroxyproline	amino acid	urea cycle; arginine-, proline-, metabolism
TRYPTOPH	tryptophan	amino acid	tryptophan metabolism
TYROSINE	tyrosine	amino acid	phenylalanine and tyrosine metabolism
UREA	urea	amino acid	urea cycle; arginine-, proline-, metabolism
V104_A	3-methyl-2-oxopentanoate	amino acid	valine, leucine and isoleucine metabolism
V167_A	S-glutathionyl-L-cysteine	amino acid	glutathione metabolism
V285_A	phenylacetylglutamine	amino acid	phenylalanine and tyrosine metabolism
V80_A	2-hydroxybutyrate	amino acid	cysteine, methionine, SAM, taurine metabolism
V82_A	2-hydroxyisobutyrate	amino acid	valine, leucine and isoleucine metabolism
VALINE	valine	amino acid	valine, leucine and isoleucine metabolism
X.2AMINOB	2-aminobutyrate	amino acid	butanoate metabolism
X.34HYDRO	4-hydroxyphenyllactate;	amino acid	phenylalanine and tyrosine metabolism
X.3INDOXY	indoxyl sulfate	amino acid	tryptophan metabolism
X.3METHYL	3-methyl-2-oxobutanoate	amino acid	valine, leucine and isoleucine metabolism
X.4ACETAM	4-acetamidobutanoate	amino acid	guanidino and acetamido metabolism
X.4METHYL	4-methyl-2-oxopentanoate	amino acid	valine, leucine and isoleucine metabolism

X.5OXOPRO	5-oxoproline	amino acid	glutathione metabolism
ARABITOL	arabitol	carbohydrate	nucleotide sugars, pentose metabolism
ERYTHRON	erythronate	carbohydrate	aminosugars metabolism
FRUCTOSE	fructose	carbohydrate	fructose, mannose, galactose, starch, and sucrose metabolism
GLUCOSE	glucose	carbohydrate	glycolysis, gluconeogenesis, pyruvate metabolism
GLYCERAT	glycerate	carbohydrate	glycolysis, gluconeogenesis, pyruvate metabolism
LACTATE	lactate	carbohydrate	glycolysis, gluconeogenesis, pyruvate metabolism
MANNOSE	mannose	carbohydrate	fructose, mannose, galactose, starch, and sucrose metabolism
PYRUVATE	pyruvate	carbohydrate	glycolysis, gluconeogenesis, pyruvate metabolism
THREITOL	threitol	carbohydrate	nucleotide sugars, pentose metabolism
X.15ANHYD	1,5-anhydro-D-glucitol	carbohydrate	glycolysis, gluconeogenesis, pyruvate metabolism
ALPHATOC	alpha-tocopherol	cofactors and vitamins	tocopherol metabolism
BILIRUBI	bilirubin (E;E)	cofactors and vitamins	hemoglobin and porphyrin metabolism
HEME	heme	cofactors and vitamins	hemoglobin and porphyrin
PANTOTHE	pantothenate	cofactors and vitamins	pantothenate and CoA metabolism
PYRIDOXA	4-pyridoxate	cofactors and vitamins	vitamin B6 metabolism
THREONAT	threonate	cofactors and vitamins	ascorbate and aldarate metabolism
V144_A	bilirubin (Z;Z)	cofactors and vitamins	hemoglobin and porphyrin metabolism
ACETYLPH	acetylphosphate	energy	oxidative phosphorylation
CITRATE	citrate	energy	krebs cycle
PHOSPHAT	phosphate	energy	oxidative phosphorylation
ADRENATE	fatty acid 22:4(7Z,10Z,13Z,16Z)	lipid	fatty acid, polyene
ANDROSTE	androsterone sulfate	lipid	sterol/steroid
ARACHIDO	fatty acid 20:4(5Z,8Z,11Z,14Z)	lipid	fatty acid, polyene
BUTYRYLC	carnitine 4:0	lipid	carnitine metabolism
CAPRATE1	fatty acid 10:0	lipid	fatty acid, saturated, even
CAPROATE	fatty acid 6:0	lipid	fatty acid, saturated, even
CAPRYLAT	fatty acid 8:0	lipid	fatty acid, saturated, even
CARNITIN	carnitine	lipid	carnitine metabolism
CHOLESTE	cholesterol	lipid	sterol/steroid
CHOLINE	choline	lipid	glycerolipid metabolism
CORTISOL	cortisol	lipid	sterol/steroid
CORTISON	cortisone	lipid	sterol/steroid
DECANOYL	carnitine 10:0	lipid	carnitine metabolism
DEHYDROI	dehydroepiandrosterone sulfate	lipid	sterol/steroid
DIHOMOLI	fatty acid 20:2(11Z,14Z)	lipid	fatty acid, polyene
DOCOSAHE	fatty acid 22:6(4Z,7Z,10Z,13Z,16Z,19Z)	lipid	fatty acid, polyene
DOCOSAPE	fatty acid 22:5(7Z,10Z,13Z,16Z,19Z)	lipid	fatty acid, polyene
EICOSAPE	fatty acid 20:5(5Z,8Z,11Z,14Z,17Z)	lipid	fatty acid, polyene
EICOSENO	fatty acid 20:1(9Z/11Z)	lipid	fatty acid, monoene

EPIANDRO	epiandrosterone sulfate	lipid	sterol/steroid
GLUTAROY	glutaroyl carnitine	lipid	carnitine metabolism
GLYCEROL	glycerol	lipid	glycerolipid metabolism
GLYCEROP	glycerophosphorylcholine	lipid	glycerolipid metabolism
HEPTANOA	fatty acid 7:0	lipid	fatty acid, saturated, odd
HEXANOYL	carnitine 6:0	lipid	carnitine metabolism
ISOBUTYR	isobutyrylcarnitine	lipid	carnitine metabolism
ISOVALER	isovalerate	lipid	fatty acid metabolism
LAURATE1	fatty acid 12:0	lipid	fatty acid, saturated, even
LINOLEAM	linoleamide 18:2(9Z,12Z)	lipid	fatty acid amide
LINOLEAT	fatty acid 18:2(9Z,12Z)	lipid	fatty acid, polyene
LINOLENA	fatty acid 18:3(n-3/n-6)	lipid	fatty acid, polyene
MARGARAT	fatty acid 17:0	lipid	fatty acid, saturated, odd
MYOINOSI	myo-inositol	lipid	inositol metabolism
MYRISTAT	fatty acid 14:0	lipid	fatty acid, saturated, even
MYRISTOL	fatty acid 14:1(9Z)	lipid	fatty acid, monoene
NONADECA	fatty acid 9:0	lipid	fatty acid, saturated, odd
OCTANOYL	carnitine 8:0	lipid	carnitine metabolism
OLEAMIDE	oleamide 18:2(9Z)	lipid	fatty acid, amide
OLEATE18	fatty acid 18:1(9Z)	lipid	fatty acid, monoene
OLEOYLCA	carnitine 18:1(9Z)	lipid	carnitine metabolism
PALMITAT	fatty acid 16:0	lipid	fatty acid, saturated, even
PALMITOL	fatty acid 16:1(9Z)	lipid	fatty acid, monoene
PALMITOY	carnitine 16:0	lipid	carnitine metabolism
PELARGON	fatty acid 9:0	lipid	fatty acid, saturated, odd
PENTADEC	fatty acid 15:0	lipid	fatty acid, saturated, odd
PROPIONY	carnitine 3:0	lipid	carnitine metabolism
STEARATE	fatty acid 18:0	lipid	fatty acid, saturated, even
STEARIDO	fatty acid 18:4(6Z,9Z,12Z,15Z)	lipid	fatty acid, polyene
THROMBOX	thromboxane B2	lipid	eicosanoid
UNDECANO	fatty acid 11:0	lipid	fatty acid, saturated, odd
V100_A	3-hydroxybutyrate	lipid	ketone bodies
V119_A	acetylcarnitine	lipid	carnitine metabolism
V173_A	fatty acid 20:3(n-3/n-6)	lipid	fatty acid, polyene
V204_A	glycerol 3-phosphate	lipid	glycerolipid metabolism
V235_A	isovalerylcarnitine	lipid	carnitine metabolism
V47_A	PE(20:4(5Z,8Z,11Z,14Z)-/0:0)	lipid	glycerolipid metabolism
V48_A	PI(20:4(5Z,8Z,11Z,14Z)-/0:0)	lipid	glycerolipid metabolism
V51_A	PC(20:3(8Z,11Z,14Z)/0:0)	lipid	glycerolipid metabolism
V54_A	PC(18:2(9Z,12Z)/0:0)	lipid	glycerolipid metabolism
V55_A	PE(18:2(9Z,12Z)/0:0)	lipid	lysolipid
V60_A	PC(18:1(9Z)/0:0)	lipid	glycerolipid metabolism
V61_A	PE(18:1(9Z))	lipid	glycerolipid metabolism
V63_A	glycerol(16:0/0:0/0:0)	lipid	monoacylglycerol
V64_A	PC(16:0/0:0)	lipid	glycerolipid metabolism
V65_A	PE(16:0/0:0)	lipid	glycerolipid metabolism
V68_A	PC(18:0/0:0)	lipid	glycerolipid metabolism
V70_A	PI(18:0/0:0)	lipid	glycerolipid metabolism
V83_A	hydroxy fatty acid 16:0	lipid	fatty acid, saturated, monohydroxy
V84_A	hydroxy fatty acid 18:0	lipid	fatty acid, saturated, monohydroxy
X.10HEPTA	fatty acid 17:1(10Z)	lipid	fatty acid, monoene, odd
X.10NONAD	fatty acid 19:1(10Z)	lipid	fatty acid, monoene, odd
X.10UNDEC	fatty acid 11:1(10Z)	lipid	fatty acid, monoene, odd
X.1ARACHI	PC(20:4(5Z,8Z,11Z,14Z)-/0:0)	lipid	glycerolipid metabolism
X.1DOCOSA	PC(22:6(4Z,7Z,10Z,13Z,-16Z,19Z)/0:0)	lipid	glycerolipid metabolism
X.1EICOSA	PC(20:2(11Z,14Z)/0:0)	lipid	lysolipid
X.1HEPTAD	PC(17:0/0:0)	lipid	glycerolipid metabolism

X.1LINOLE	glycerol(18:2(9Z,12Z)/0:0-/0:0)	lipid	monoacylglycerol
X.1MYRIST	PC(14:0/0:0)	lipid	glycerolipid metabolism
X.1OLEOYL	glycerol(18:1(9Z)/0:0/0:0)	lipid	monoacylglycerol
X.1PALMIT	PC(16:1(9Z)/0:0)	lipid	glycerolipid metabolism
X.2LINOLE	PC(0:0/18:2(9Z,12Z))	lipid	glycerolipid metabolism
X.2METHYL	2-methylbutyroylcarnitine	lipid	carnitine metabolism
X.2OLEOYL	PC(0:0/18:1(9Z))	lipid	glycerolipid metabolism
X.2PALMIT	PC(0:0/16:0)	lipid	glycerolipid metabolism
X.2STEARO	PC(0:0/18:0)	lipid	glycerolipid metabolism
X.2TETRAD	2-tetradecenoyl carnitine	lipid	carnitine metabolism
X.3CARBOX	3-carboxy-4-methyl-5-propyl-2-furanpropanoate	lipid	fatty acid, furan, dicarboxylate
X.3DEHYDR	3-dehydrocarnitine	lipid	carnitine metabolism
X.5DODECE	fatty acid 12:0(5Z)	lipid	fatty acid, monoene
X.7ALPHAH	7-hoca	lipid	sterol/steroid
GUANOSIN	guanosine	nucleotide	purine metabolism, guanine containing
HYPOXANT	hypoxanthine	nucleotide	purine metabolism, (hypo)xanthine/inosine containing
INOSINE	inosine	nucleotide	purine metabolism, (hypo)xanthine/inosine containing
N1METHYL	1-methyladenosine	nucleotide	purine metabolism, adenine containing
PSEUDOUR	pseudouridine	nucleotide	pyrimidine metabolism, uracil containing
URATE	urate	nucleotide	purine metabolism, urate metabolism
URIDINE	uridine	nucleotide	pyrimidine metabolism, uracil containing
X.7METHYL	7-methylguanine	nucleotide	purine metabolism, guanine containing
XANTHINE	xanthine	nucleotide	purine metabolism, (hypo)xanthine/inosine containing
ADPSGEGD	ADpSGEGDFXAEGGGVR	peptide	fibrinogen cleavage peptide
ADSGEGDF	ADSGEGDFXAEGGGVR	peptide	fibrinogen cleavage peptide
ASPARTYL	aspartyl-phenylalanine	peptide	dipeptide
DSGEGDFX	DSGEGDFXAEGGGVR	peptide	fibrinogen cleavage peptide
HWESASXX	HWESASXX	peptide	polypeptide
PROHYDRO	pro-hydroxy-pro	peptide	dipeptide
V188_A	gamma-glutamyl-glutamine	peptide	g-glutamyl
V190_A	gamma-glutamyl-leucine	peptide	g-glutamyl
V192_A	gamma-glutamyl-phenylalanine	peptide	g-glutamyl
V194_A	gamma-glutamyl-tyrosine	peptide	g-glutamyl
V195_A	gamma-glutamyl-valine	peptide	dipeptide
BENZOATE	benzoate	xenobiotics	benzoate metabolism
CAFFEINE	caffeine	xenobiotics	xanthine metabolism
CATECHOL	catecholsulfate	xenobiotics	benzoate metabolism
ERYTHRIT	erythritol	xenobiotics	sugar, sugar substitute, starch
HIPPURAT	hippurate	xenobiotics	benzoate metabolism
PARAXANT	paraxanthine	xenobiotics	xanthine metabolism
PHENOLSU	phenylsulfate	xenobiotics	chemical
PIPERINE	piperine	xenobiotics	food component/plant
THEOBROM	theobromine	xenobiotics	xanthine metabolism
THEOPHYL	theophylline	xenobiotics	xanthine metabolism
X.4VINYLP	4-vinylphenylsulfate	xenobiotics	benzoate metabolism

Table C.3.: **GGM-defined phenotype sets on Biocrates metabolites.** Phenotype sets estimated with the GGM on Biocrates metabolites in KORA F4. These sets were analyzed for enrichment with PSEA in KORA F4. Replication was performed in TwinsUK data.

set name	elements of the metabolite set
set_1_b	C0, C3
set_2_b	C10, C10.1, C10.2, C12, C12.1, C14, C14.1, C14.2, C16, C18, C18.1, C18.2, C2, C4.1.DC...C6, C8
set_3_b	C14.1.OH, C7.DC
set_4_b	C5.M.DC, C5.1.DC, C6.1
set_5_b	Gly, Ser, Thr
set_6_b	Met, Trp, Tyr
set_7_b	Val, xLeu
set_8_b	PC.aa.C24.0, PC.aa.C26.0, lysoPC.a.C24.0, lysoPC.a.C26.1, lysoPC.a.C28.1
set_9_b	PC.aa.C28.1, PC.ae.C34.1, PC.ae.C36.1, SM..OH..C14.1, SM..OH..C16.1, SM..OH..C22.1, SM..OH..C22.2, SM..OH..C24.1, SM.C16.0, SM.C16.1, SM.C18.0, SM.C18.1, SM.C24.0, SM.C24.1, SM.C26.0, SM.C26.1
set_10_b	PC.aa.C30.0, PC.aa.C32.0, PC.ae.C30.0
set_11_b	PC.aa.C32.1, PC.aa.C34.1, PC.aa.C34.2, PC.aa.C36.1, PC.aa.C36.2, PC.aa.C36.3, PC.aa.C36.4, PC.aa.C36.5, PC.aa.C38.3, PC.aa.C38.4, PC.aa.C38.5, PC.aa.C38.6, PC.aa.C40.4, PC.aa.C40.5, PC.aa.C40.6, PC.ae.C38.3, PC.ae.C40.0, PC.ae.C40.6, PC.ae.C42.0, lysoPC.a.C16.0, lysoPC.a.C16.1, lysoPC.a.C17.0, lysoPC.a.C18.0, lysoPC.a.C18.1, lysoPC.a.C18.2, lysoPC.a.C20.3, lysoPC.a.C20.4
set_12_b	PC.aa.C32.2, PC.aa.C34.4, PC.aa.C36.6, PC.ae.C38.0, PC.ae.C40.1, PC.ae.C40.2, PC.ae.C42.2, PC.ae.C42.3
set_13_b	PC.aa.C32.3, PC.aa.C34.3
set_14_b	PC.aa.C36.0, PC.aa.C38.0, PC.aa.C42.0, PC.aa.C42.1, PC.ae.C34.2, PC.ae.C34.3, PC.ae.C36.3, PC.ae.C36.4, PC.ae.C36.5, PC.ae.C38.4, PC.ae.C38.5, PC.ae.C38.6, PC.ae.C40.4, PC.ae.C40.5, PC.ae.C42.4, PC.ae.C42.5, PC.ae.C44.4, PC.ae.C44.5, PC.ae.C44.6
set_15_b	PC.aa.C40.2, PC.aa.C40.3, PC.aa.C42.5, PC.aa.C42.6
set_16_b	PC.ae.C32.1, PC.ae.C32.2
set_17_b	C10, C8
set_18_b	C18.1, C18.2
set_19_b	PC.aa.C24.0, PC.aa.C26.0
set_20_b	PC.aa.C30.0, PC.aa.C32.0
set_21_b	PC.aa.C32.1, lysoPC.a.C16.1
set_22_b	PC.aa.C34.2, PC.aa.C36.2
set_23_b	PC.aa.C36.4, PC.aa.C38.4
set_24_b	PC.aa.C36.5, PC.aa.C38.5, PC.aa.C40.4, PC.aa.C40.5
set_25_b	PC.aa.C36.6, PC.ae.C38.0
set_26_b	PC.aa.C38.0, PC.ae.C38.6
set_27_b	PC.aa.C38.6, PC.aa.C40.6
set_28_b	PC.aa.C40.2, PC.aa.C40.3
set_29_b	PC.ae.C34.2, PC.ae.C36.3
set_30_b	PC.ae.C34.3, PC.ae.C36.5
set_31_b	PC.ae.C36.4, PC.ae.C38.5
set_32_b	PC.ae.C38.4, PC.ae.C40.4, PC.ae.C42.4
set_33_b	PC.ae.C42.5, PC.ae.C44.4, PC.ae.C44.5
set_34_b	lysoPC.a.C16.0, lysoPC.a.C18.0
set_35_b	lysoPC.a.C20.3, lysoPC.a.C20.4
set_36_b	SM..OH..C14.1, SM..OH..C16.1
set_37_b	SM..OH..C22.1, SM..OH..C22.2, SM.C24.0, SM.C24.1
set_38_b	SM.C16.0, SM.C16.1, SM.C18.0, SM.C18.1

Table C.4.: **GGM-defined phenotype sets on Metabolon metabolites.** Phenotype sets estimated with the GGM on Metabolon metabolites in KORA F4. These sets were analyzed for enrichment with PSEA in KORA F4. For the replication in TwinsUK two metabolites were not available, these and the corresponding sets are marked with stars.

set name	elements of the metabolite set
set_1_m	X.1ARACHI, V47_A, X.1DOCOSA, X.1EICOSA, V51_A, V54_A, V55_A, V60_A, V64_A, V68_A, X.2LINOLE, X.2OLEOYL, X.2PALMIT, X.2STEARO, X.3CARBOX, ARACHIDO, V173_A, DOCOSAHE, *LINOLEAM*, *OLEAMIDE*
set_2_m	V48_A, V70_A
set_3_m	X.1LINOLE, X.1OLEOYL
set_4_m	X.1MYRIST, X.1PALMIT
set_5_m	V61_A, V65_A
set_6_m	X.10HEPTA, X.10NONAD, X.5DODECE, MARGARAT, MYRISTOL, PALMITOL
set_7_m	X.2AMINOB, V80_A
set_8_m	V83_A, V84_A
set_9_m	X.2TETRAD, DECANOYL, HEXANOYL, OCTANOYL
set_10_m	X.34HYDRO, ALPHAHYD
set_11_m	V100_A, ALANINE
set_12_m	X.3INDOXY, PCRESOLS, V285_A
set_13_m	X.3METHYL, V104_A, X.4METHYL, ISOLEUCI, LEUCINE, VALINE
set_14_m	V119_A, PROPIONY
set_15_m	ACETYLPH, CHOLESTE, PHOSPHAT
set_16_m	ADSGEGDF, DSGEGDFX
set_17_m	ANDROSTE, EPIANDRO
set_18_m	ASPARTAT, CYSTEINE
set_19_m	BETAINE, CHOLINE
set_20_m	BILIRUBI, V144_A
set_21_m	CAFFEINE, PARAXANT, THEOPHYL
set_22_m	CAPRATE1, CAPROATE, CAPRYLAT, HEPTANOA, PELARGON
set_23_m	CATECHOL, HIPPURAT
set_24_m	CORTISOL, CORTISON
set_25_m	CREATINE, PYROGLUT
set_26_m	DEHYDROI, THROMBOX
set_27_m	DIHOMOLI, EICOSENO, LINOLEAT, LINOLENA, STEARIDO
set_28_m	DOCOSAPE, EICOSAPE
set_29_m	V188_A, GLUTAMIN
set_30_m	V194_A, TYROSINE
set_31_m	GLUCOSE, MANNOSE
set_32_m	GLYCERAT, THREONAT
set_33_m	GLYCINE, SERINE
set_34_m	GUANOSIN, INOSINE
set_35_m	ISOVALER, V235_A
set_36_m	LAURATE1, MYRISTAT
set_37_m	OLEATE18, PALMITAT
set_38_m	OLEOYLCA, PALMITOY
set_39_m	PANTOTHE, PYRIDOXA
set_40_m	X.1ARACHI, V47_A, X.1DOCOSA, V55_A, DOCOSAHE
set_41_m	V51_A, V173_A
set_42_m	V54_A, X.2PALMIT
set_43_m	X.10HEPTA, PALMITOL
set_44_m	X.2STEARO, *LINOLEAM*, *OLEAMIDE*
set_45_m	V104_A, X.4METHYL, ISOLEUCI, LEUCINE
set_46_m	CHOLESTE, PHOSPHAT
set_47_m	DECANOYL, HEXANOYL, OCTANOYL
set_48_m	HEPTANOA, PELARGON
set_49_m	PCRESOLS, V285_A
set_50_m	PARAXANT, THEOPHYL

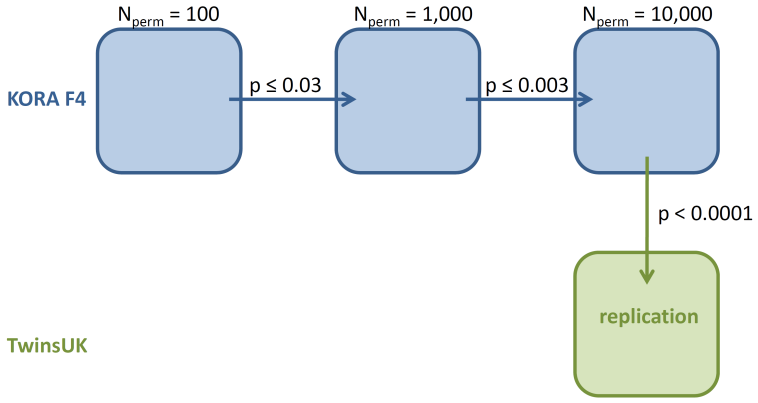


Figure C.1.: **Permutation scheme of PSEA for predefined metabolite sets.** This figure visualizes the stepwise permutation scheme and replication that was used in application of PSEA on predefined metabolite sets. For both metabolite panels, Biocrates and Metabolon, the same permutation strategy was applied.

Table C.5.: **PSEA results on Metabolon metabolites for sets differing in KORA F4 and TwinsUK.** For the presented genes the given phenotype sets showed an enrichment with a p-value below 10^{-4} in PSEA on KORA F4 and were not replicable in TwinsUK data. The phenotype *set_1_m* had to be modified for replication in TwinsUK and it was not possible to test *set_44_m* in TwinsUK. Gene groups are separated with background color.

chr.	start pos.	stop pos.	phenotype sets (set_*_m)
3	15143274	15467811	44
genes	COL6A4P1, SH3BP5		
8	87838015	88501537	44
genes	CNBD1		
11	61094821	61527475	1
genes	FEN1, C11orf9, DKFZP434K028, MIR1908, MIR611, C11orf10, FADS1, DAGLA, FADS2, BEST1, FADS3		

D. Supplementary Information: New Identfied Phenotype Sets

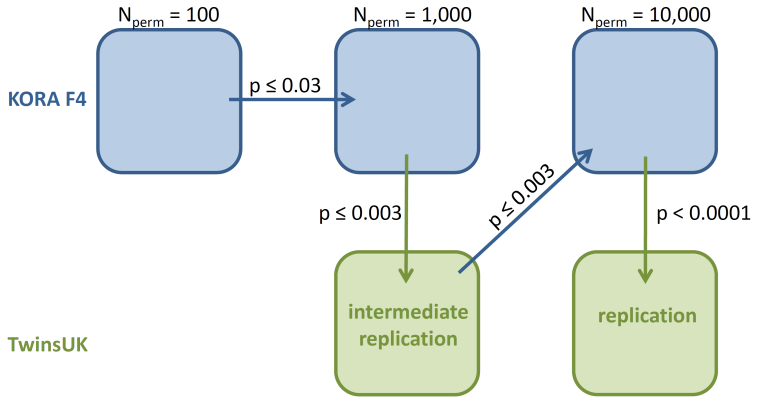


Figure D.1.: **Permutation scheme of PSEA for new identifed metabolite sets.** This figure visualizes the stepwise permutation scheme with intermediate replication that was used in application of PSEA with the extension of new identfication of metabolite sets. For both metabolite panels, Biocrates and Metabolon, the same permutation strategy was applied.

D.1. New Identified Metabolite Sets

Table D.1.: **New phenotype sets on Biocrates metabolites.** All new identified phenotype sets that were signlificantly enriched in KORA F4 and replicated in TwinsUK. num: number of all metabolites in the metabolite set.

gene	P (KORA F4)	P (TwinsUK)	num	elements of the phenotype set
MFSD2A	<0.0001	0.0010	7	lysoPC.a.C16.0 lysoPC.a.C17.0 lysoPC.a.C18.0 lysoPC.a.C18.1 lysoPC.a.C20.4
MYCL1	<0.0001	0.0010	7	lysoPC.a.C16.0 lysoPC.a.C17.0 lysoPC.a.C18.0 lysoPC.a.C18.1 lysoPC.a.C20.4
MSH4	<0.0001	0.0001	12	C10 C10.1 C10.2 C12.DC C16.2.OH C4.1.DC...C6 C5.1.DC C8 PC.aa.C38.0 PC.ae.C38.6
RABGGTB	<0.0001	0.0002	11	C10 C10.1 C10.2 C12.DC C4.1.DC...C6 C5.1.DC C8 PC.aa.C38.0 PC.ae.C38.6
SNORD45A	<0.0001	0.0002	11	C10 C10.1 C10.2 C12.DC C4.1.DC...C6 C5.1.DC C8 PC.aa.C38.0 PC.ae.C38.6
SNORD45B	<0.0001	0.0002	11	C10 C10.1 C10.2 C12.DC C4.1.DC...C6 C5.1.DC C8 PC.aa.C38.0 PC.ae.C38.6
SNORD45C	<0.0001	0.0002	11	C10 C10.1 C10.2 C12.DC C4.1.DC...C6 C5.1.DC C8 PC.aa.C38.0 PC.ae.C38.6
DKFZp686O1327	<0.0001	0.0003	13	C0 C16.2.OH C18 Arg Thr PC.aa.C36.5 PC.aa.C36.6 PC.aa.C40.3 PC.ae.C36.5 lysoPC.a.C18.1 lysoPC.a.C20.4
CPS1	<0.0001	<0.0001	12	C3.DC.M...C5.OH C5.M.DC C5.1 C9 Gly Ser PC.ae.C36.5 PC.ae.C42.3 lysoPC.a.C18.0 SM..OH..C24.1
MYL1	<0.0001	<0.0001	4	C9 SM..OH..C24.1
PDCD6IP	<0.0001	0.0001	14	C10 C10.1 C12.1 C14.1 C16.1.OH C18.1 C4.1.DC...C6 C7.DC C8 lysoPC.a.C16.0 lysoPC.a.C18.0 lysoPC.a.C18.2
C4orf46	<0.0001	<0.0001	5	C10 C8 SM.C26.0
PPID	<0.0001	<0.0001	5	C10 C8 SM.C26.0
ETFDH	<0.0001	0.0001	5	C10 C8 SM.C26.0
FNIP2	<0.0001	0.0001	7	C10 C8 lysoPC.a.C17.0 lysoPC.a.C18.0 SM.C26.0
IL3	<0.0001	0.0006	10	C16 C18.1 C18.2 C3 C5 C7.DC PC.ae.C40.5 SM..OH..C14.1
PDLIM4	<0.0001	<0.0001	15	C0 C10.1 C16 C16.1 C18.1 C18.2 C2 C3 C5 C7.DC PC.ae.C40.5 PC.ae.C42.5 SM.C24.1
P4HA2	<0.0001	0.0001	22	C0 C10.1 C10.2 C16 C16.1 C18.1 C18.2 C2 C3 C4 C5 C7.DC PC.ae.C34.1 PC.ae.C40.5 PC.ae.C42.5 PC.ae.C44.4 SM..OH..C16.1 SM..OH..C22.2 SM.C18.0 SM.C24.1
SLC22A4	<0.0001	0.0001	24	C0 C10.1 C10.2 C16 C16.1 C16.1.OH C18.1 C18.2 C2 C3 C4 C5 C7.DC PC.ae.C34.1 PC.ae.C40.5 PC.ae.C42.5 PC.ae.C44.4 SM..OH..C14.1 SM..OH..C16.1 SM..OH..C22.2 SM.C18.0 SM.C24.1
SLC22A5	<0.0001	0.0003	24	C0 C10.2 C14 C16 C16.1 C16.1.OH C18.1 C18.2 C2 C3 C4 C4.1.DC...C6 C5 PC.ae.C34.1 PC.ae.C40.5 PC.ae.C42.5 PC.ae.C44.4 SM..OH..C14.1 SM..OH..C16.1 SM..OH..C22.2 SM.C18.0 SM.C24.1
IGF2R	<0.0001	<0.0001	13	C14.1 C16.1.OH C2 C3 C4 C8.1 Gln PC.aa.C40.3 PC.ae.C40.0 PC.ae.C40.2 lysoPC.a.C26.1
SLC22A1	<0.0001	<0.0001	14	C14.1 C16.1.OH C2 C3 C4 C8.1 Gln PC.aa.C40.3 PC.ae.C38.2 PC.ae.C40.0 PC.ae.C40.2 lysoPC.a.C26.1
INTS8	<0.0001	0.0008	5	C10.2 PC.aa.C36.1 PC.aa.C42.2
SLC16A9	<0.0001	0.0011	17	C0 C2 C3 C3.DC...C4.OH C4 C4.1.DC...C6 C8.1 PC.aa.C32.3 PC.ae.C40.5 lysoPC.a.C16.0 lysoPC.a.C20.3 lysoPC.a.C20.4 SM.C24.0 SM.C24.1 H1
BEST1	<0.0001	<0.0001	46	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.1 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.5 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.4 SM.C20.2

C11orf10	<0.0001	<0.0001	60	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC Orn PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.1 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 lysoPC.a.C28.1 SM.C20.2
C11orf9	<0.0001	<0.0001	57	C5.DC...C6.OH C5.1.DC PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.1 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 lysoPC.a.C28.1 SM.C20.2
DAGLA	<0.0001	<0.0001	54	PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.1 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C18.0 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 lysoPC.a.C28.1 SM.C20.2
DKFZP434K028	<0.0001	<0.0001	60	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC Orn PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.1 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 lysoPC.a.C28.1 SM.C20.2
FADS1	<0.0001	<0.0001	54	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC Orn PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 SM.C20.2

FADS2	<0.0001	60	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC Orn PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.1 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C38.2 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.2 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C40.6 PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 lysoPC.a.C28.1 SM.C20.2
FADS3	<0.0001	47	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.1 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.5 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.4 SM.C20.2
FEN1	<0.0001	58	C5.DC...C6.OH C5.1.DC Orn PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.1 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C40.6 PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 lysoPC.a.C28.1 SM.C20.2
MIR1908	<0.0001	52	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC Orn PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 SM.C20.2
MIR611	<0.0001	54	C3 C3.DC.M...C5.OH C5.DC...C6.OH C5.1.DC Orn PC.aa.C24.0 PC.aa.C32.0 PC.aa.C32.2 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.1 lysoPC.a.C28.0 SM.C20.2
RAB3IL1	<0.0001	34	C5.1.DC PC.aa.C24.0 PC.aa.C32.0 PC.aa.C34.4 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C40.2 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C42.1 PC.aa.C42.4 PC.aa.C42.6 PC.ae.C34.2 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C40.1 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C42.1 PC.ae.C42.5 PC.ae.C44.5 PC.ae.C44.6 lysoPC.a.C20.4
C12orf75	<0.0001	0.0010	14 C10.2 C14.2.OH C3.DC...C4.OH C4.1.DC...C6 C5 PC.aa.C36.2 PC.ae.C40.2 SM..OH..C14.1 SM..OH..C16.1 SM..OH..C22.1 SM..OH..C22.2 SM..OH..C24.1
ACADS	<0.0001	<0.0001	5 C3 C4 PC.aa.C36.2
CABP1	<0.0001	<0.0001	5 C3 C4 PC.aa.C36.2

MLEC	<0.0001	<0.0001	5	C3 C4 PC.aa.C36.2
POP5	<0.0001	<0.0001	5	C3 C4 PC.aa.C36.2
SPPL3	<0.0001	<0.0001	13	C4 Orn PC.aa.C38.0 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.ae.C38.6 PC.ae.C40.0 PC.ae.C40.6 SM..OH..C24.1
UNC119B	<0.0001	<0.0001	5	C3 C4 PC.aa.C36.2
OASL	<0.0001	0.0001	6	C4 PC.aa.C42.1 PC.ae.C40.6 SM..OH..C24.1
SGPP1	<0.0001	<0.0001	5	PC.aa.C28.1 PC.ae.C30.2 SM..OH..C14.1
SYNE2	<0.0001	0.0004	11	C12 C12.1 C14.1 C14.2 C18.2 C2 PC.aa.C28.1 PC.ae.C30.2 SM..OH..C14.1

Table D.2.: **New phenotype sets on Metabolon metabolites.** All new identified phenotype sets that were signlificantly enriched in KORA F4 and replicated in TwinsUK. num: number of all metabolites in the metabolite set.

gene	P (KORA F4)	P (TwinsUK)	num	elements of the phenotype set
CYP4A11	<0.0001	<0.0001	3	X.1MYRIST X.1PALMIT X.10UNDEC
CYP4B1	<0.0001	<0.0001	6	X.10UNDEC V194_A GLUTAROY HEME PROPIONY TYROSINE
CYP4X1	<0.0001	<0.0001	3	X.1MYRIST X.1PALMIT X.10UNDEC
CYP4Z2P	<0.0001	<0.0001	3	X.1PALMIT X.10UNDEC GLUTAROY
KIAA0494	<0.0001	<0.0001	5	X.10UNDEC V194_A HEME PROPIONY TYROSINE
DIRAS3	<0.0001	0.0001	4	V61_A V104_A GLYCERAT GLYCEROL
ACADM	<0.0001	<0.0001	6	CAPRATE1 DECANOYL ERYTHRIT HEXANOYL OCTANOYL PENTADEC
MSH4	<0.0001	<0.0001	6	CAPRATE1 DECANOYL ERYTHRIT HEXANOYL OCTANOYL PENTADEC
RABGGTB	<0.0001	<0.0001	6	CAPRATE1 DECANOYL ERYTHRIT HEXANOYL OCTANOYL PENTADEC
SLC44A5	<0.0001	<0.0001	16	V47_A V48_A X.1MYRIST V70_A ADRENATE BETAINE DECANOYL DOCOSAPE HEXANOYL ISOLEUCI LEUCINE NACETYLA OCTANOYL PHOSPHAT PROHYDRO TYROSINE
SNORD45A	<0.0001	<0.0001	6	CAPRATE1 DECANOYL ERYTHRIT HEXANOYL OCTANOYL PENTADEC
SNORD45B	<0.0001	<0.0001	6	CAPRATE1 DECANOYL ERYTHRIT HEXANOYL OCTANOYL PENTADEC
SNORD45C	<0.0001	<0.0001	6	CAPRATE1 DECANOYL ERYTHRIT HEXANOYL OCTANOYL PENTADEC
ST6GALNAC3	<0.0001	<0.0001	11	CGLYCOSY DECANOYL DOCOSAPE V188_A GLUTAMIN HEME HEXANOYL INDOLEAC OCTANOYL PROHYDRO VALINE
ASB17	<0.0001	0.0001	5	DECANOYL HEME HEXANOYL OCTANOYL VALINE
C2orf16	<0.0001	<0.0001	17	V51_A V55_A X.1OLEOYL V60_A V63_A V80_A X.2OLEOYL X.3METHYL V104_A ASPARAGI ERYTHRIT V188_A GLUTAMIN LACTATE MANNOSE PANTOTHE THREONIN
FNDC4	<0.0001	<0.0001	16	V51_A V55_A X.1OLEOYL V60_A V63_A V80_A X.2OLEOYL V104_A ASPARAGI ERYTHRIT V188_A GLUTAMIN LACTATE MANNOSE PANTOTHE THREONIN
GCKR	<0.0001	<0.0001	16	V51_A V55_A X.1OLEOYL V60_A V63_A V80_A V104_A ASPARAGI ERYTHRIT V188_A GLUTAMIN HISTIDIN LACTATE MANNOSE PANTOTHE THREONIN
IFT172	<0.0001	<0.0001	17	V51_A V55_A X.1OLEOYL V60_A V63_A V80_A X.2OLEOYL V104_A ASPARAGI ERYTHRIT V188_A GLUTAMIN HISTIDIN LACTATE MANNOSE PANTOTHE THREONIN
PPM1G	<0.0001	<0.0001	12	V55_A X.1OLEOYL V63_A V80_A V104_A ASPARAGI V188_A GLUTAMIN HISTIDIN LACTATE MANNOSE PANTOTHE
SLC30A3	<0.0001	0.0003	4	V80_A MANNOSE PANTOTHE PYROGLUT
CEP68	<0.0001	<0.0001	3	X.2AMINOB ALANINE THREONIN
SLC1A4	<0.0001	<0.0001	5	X.2AMINOB X.7ALPHAH ALANINE GLUTAMIN THREONIN
ALMS1	<0.0001	<0.0001	6	X.1LINOLE X.2PALMIT ASPARTAT CYSTEINE HEME NACETYLO
ALMS1P	<0.0001	<0.0001	5	X.2PALMIT ASPARTAT CYSTEINE HEME NACETYLO
C2orf7	<0.0001	<0.0001	5	X.1LINOLE ASPARTAT NACETYLO PHENYLAL VALINE
EGR4	<0.0001	<0.0001	4	X.1LINOLE ASPARTAT NACETYLO PHENYLAL
FBXO41	<0.0001	<0.0001	5	X.1LINOLE ASPARTAT NACETYLO PHENYLAL VALINE

NAT8	<0.0001	<0.0001	5	X.2PALMIT ASPARTAT CYSTEINE HEME NACETYLO
NAT8B	<0.0001	<0.0001	3	ASPARTAT CYSTEINE NACETYLO
CPS1	<0.0001	<0.0001	8	X.5DODECE BETAINE CREATINE ERYTHRON GLYCINE PYROGLUT SERINE UNDECANO
DNAJB3	<0.0001	<0.0001	6	X.2TETRAD X.5OXOPRO BILIRUBI V144_A MYRISTOL SEROTONI
UGT1A1	<0.0001	<0.0001	8	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL SEROTONI
UGT1A10	<0.0001	<0.0001	9	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL ORNITHIN SEROTONI
UGT1A3	<0.0001	<0.0001	8	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL SEROTONI
UGT1A4	<0.0001	<0.0001	8	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL SEROTONI
UGT1A5	<0.0001	<0.0001	8	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL SEROTONI
UGT1A6	<0.0001	<0.0001	8	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL SEROTONI
UGT1A7	<0.0001	<0.0001	8	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL SEROTONI
UGT1A8	<0.0001	<0.0001	9	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL ORNITHIN SEROTONI
UGT1A9	<0.0001	<0.0001	8	X.2TETRAD X.5OXOPRO BILIRUBI V144_A LINOLEAT LINOLENA MYRISTOL SEROTONI
USP40	<0.0001	0.0002	4	V70_A BILIRUBI V144_A ORNITHIN
MIR138-1	<0.0001	0.0004	5	V51_A V54_A X.2OLEOYL CREATINI V194_A
SLC2A9	<0.0001	0.0002	13	X.3CARBOX X.7METHYL GLYCEROL HEXANOYL HISTIDIN ISOVALER MYRISTOL NACETYLO OLEATE18 PALMITOL PYRIDOXA URATE XANTHINE
AP4M1	<0.0001	<0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
CNPY4	<0.0001	<0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
CYP3A4	<0.0001	<0.0001	4	X.1OLEOYL ANDROSTE EPIANDRO V285_A
CYP3A43	<0.0001	<0.0001	4	X.1OLEOYL ANDROSTE EPIANDRO V285_A
CYP3A5	<0.0001	<0.0001	3	ANDROSTE EPIANDRO V285_A
CYP3A7	<0.0001	<0.0001	3	ANDROSTE EPIANDRO V285_A
FAM200A	<0.0001	<0.0001	3	ANDROSTE EPIANDRO GLUTAROY
GATS	<0.0001	<0.0001	4	V119_A ANDROSTE EPIANDRO PROPIONY
MBLAC1	<0.0001	<0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
OR2AE1	<0.0001	<0.0001	3	X.1OLEOYL ANDROSTE EPIANDRO
PVRIG	<0.0001	<0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
SPDYE3	<0.0001	<0.0001	4	V119_A ANDROSTE EPIANDRO PROPIONY
STAG3	<0.0001	<0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
TAF6	<0.0001	<0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
TRIM4	<0.0001	<0.0001	3	X.1OLEOYL ANDROSTE EPIANDRO
ZNF3	<0.0001	<0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
ZNF394	<0.0001	<0.0001	3	ANDROSTE EPIANDRO GLUTAROY
ZNF498	<0.0001	<0.0001	3	ANDROSTE EPIANDRO GLUTAROY
ZNF655	<0.0001	<0.0001	3	ANDROSTE EPIANDRO GLUTAROY
C7orf59	<0.0001	0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
COPS6	<0.0001	0.0001	4	ANDROSTE EPIANDRO PROPIONY URATE
MCM7	<0.0001	0.0002	4	ANDROSTE EPIANDRO PROPIONY URATE
MIR106B	<0.0001	0.0002	4	ANDROSTE EPIANDRO PROPIONY URATE
MIR25	<0.0001	0.0002	4	ANDROSTE EPIANDRO PROPIONY URATE
MIR93	<0.0001	0.0002	4	ANDROSTE EPIANDRO PROPIONY URATE
HEATR7A	<0.0001	<0.0001	2	X.2LINOLE X.5OXOPRO

SCXA	<0.0001	<0.0001	2	X.2LINOLE X.5OXOPRO
SCXB	<0.0001	<0.0001	2	X.2LINOLE X.5OXOPRO
MIR661	<0.0001	0.0002	4	V60_A V64_A X.2PALMIT X.5OXOPRO
PARP10	<0.0001	0.0003	4	V60_A V64_A X.2PALMIT X.5OXOPRO
LINGO2	<0.0001	<0.0001	18	X.1LINOLE V55_A X.1OLEOYL X.10UNDEC X.2TETRAD ARACHIDO ASPARTAT BETAINE CREATINE V167_A V173_A DSGEGDFX GLUTAMAT V235_A METHIONI PHENOLSU PYROGLUT XANTHINE
OBP2B	<0.0001	0.0002	7	X.3CARBOX ADPSGEGD ADSGEGDF ASPARTYL CATECHOL DSGEGDFX UREA
SLC16A9	<0.0001	<0.0001	8	ASPARTYL BUTYRYLC CARNITIN CHOLESTE V173_A GLUTAMAT ISOVALER PROPIONY
C11orf10	<0.0001	<0.0001	21	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ARACHIDO ASPARAGI V173_A DOCOSAPE EICOSAPE ISOLEUCI LINOLEAM PHOSPHAT PROPIONY STEARIDO URATE
C11orf9	<0.0001	<0.0001	17	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ARACHIDO ARGININE V173_A DOCOSAPE EICOSAPE ISOLEUCI STEARIDO
DKFZP434K028	<0.0001	<0.0001	19	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ARACHIDO V173_A DOCOSAPE EICOSAPE ISOLEUCI LINOLEAM PROPIONY STEARIDO URATE
FADS1	<0.0001	<0.0001	22	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ALANINE ARACHIDO ASPARAGI V173_A DOCOSAPE EICOSAPE ISOLEUCI LINOLEAM PHOSPHAT PROPIONY STEARIDO URATE
FADS2	<0.0001	<0.0001	21	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ARACHIDO ASPARAGI V173_A DOCOSAPE EICOSAPE ISOLEUCI LINOLEAM PHOSPHAT PROPIONY STEARIDO URATE
FEN1	<0.0001	<0.0001	17	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ARACHIDO ARGININE V173_A DOCOSAPE EICOSAPE ISOLEUCI STEARIDO
MIR1908	<0.0001	<0.0001	22	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ALANINE ARACHIDO ASPARAGI V173_A DOCOSAPE EICOSAPE ISOLEUCI LINOLEAM PHOSPHAT PROPIONY STEARIDO URATE
MIR611	<0.0001	<0.0001	21	X.1ARACHI V47_A V48_A X.1EICOSA V51_A X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ARACHIDO ASPARAGI V173_A DOCOSAPE EICOSAPE ISOLEUCI LINOLEAM PHOSPHAT PROPIONY STEARIDO URATE
DAGLA	<0.0001	0.0001	16	X.1ARACHI V47_A V48_A X.1EICOSA X.1LINOLE V54_A V55_A X.2LINOLE ADRENATE ARACHIDO ARGININE V173_A DOCOSAPE EICOSAPE ISOLEUCI STEARIDO
FADS3	<0.0001	0.0003	20	X.1ARACHI V47_A V48_A X.1EICOSA V51_A V54_A V55_A X.2LINOLE ADRENATE ALANINE ARACHIDO ASPARAGI V173_A DOCOSAPE EICOSAPE LINOLEAM PHOSPHAT PROPIONY STEARIDO URATE
TIMELESS	<0.0001	<0.0001	9	ALANINE V188_A GLUTAMIN LYSINE METHIONI PIPECOLA THREONIN TYROSINE XANTHINE
MIP	<0.0001	0.0001	9	ALANINE V188_A GLUTAMIN LYSINE METHIONI PIPECOLA THREONIN TYROSINE XANTHINE
GLS2	<0.0001	0.0002	9	ALANINE V188_A GLUTAMIN LYSINE METHIONI PIPECOLA THREONIN TYROSINE XANTHINE
APOF	<0.0001	0.0003	8	ALANINE V188_A GLUTAMIN METHIONI PIPECOLA THREONIN TYROSINE XANTHINE
ACADS	<0.0001	<0.0001	8	BUTYRYLC CITRULLI DIMETHYL GLYCEROL HWESASXX PALMITOL PROHYDRO SEROTONI
C12orf43	<0.0001	<0.0001	6	X.4VINYLP BUTYRYLC CYSTEINE V235_A PARAXANT THEOPHYL
CABP1	<0.0001	<0.0001	5	X.1HEPTAD V65_A BUTYRYLC GLYCEROL URATE
COQ5	<0.0001	<0.0001	5	X.1HEPTAD V65_A BUTYRYLC GLYCEROL URATE
HNF1A	<0.0001	<0.0001	7	X.4VINYLP BUTYRYLC CITRULLI V235_A PARAXANT SEROTONI THEOPHYL

HPD	<0.0001	<0.0001	3	X.1ARACHI V82_A PYRIDOXA
MLEC	<0.0001	<0.0001	5	X.1HEPTAD V65_A BUTYRYLC GLYCEROL URATE
NCRNA00262	<0.0001	<0.0001	7	X.4VINYLP BUTYRYLC CYSTEINE V235_A PARAXANT SEROTONI THEOPHYL
OASL	<0.0001	<0.0001	6	X.4VINYLP BUTYRYLC CYSTEINE V235_A PARAXANT THEOPHYL
P2RX7	<0.0001	<0.0001	6	ADPSGEGD BUTYRYLC CITRULLI CYSTEINE FRUCTOSE HYPOXANT
POP5	<0.0001	<0.0001	5	X.1HEPTAD V65_A BUTYRYLC GLYCEROL URATE
PSMD9	<0.0001	<0.0001	3	X.1ARACHI V82_A PYRIDOXA
RNF10	<0.0001	<0.0001	6	X.1HEPTAD V65_A X.4VINYLP BUTYRYLC LINOLEAT URATE
SPPL3	<0.0001	<0.0001	11	X.4VINYLP BUTYRYLC CITRULLI DIMETHYL HWESASXX V235_A PALMITOL PARAXANT PROHYDRO SEROTONI THEOPHYL
UNC119B	<0.0001	<0.0001	9	X.1HEPTAD BUTYRYLC CITRULLI DIMETHYL GLYCEROL HWESASXX PROHYDRO SEROTONI URATE
WDR66	<0.0001	<0.0001	3	X.1ARACHI V82_A PYRIDOXA
UBL3	<0.0001	0.0001	10	V83_A V84_A X.3DEHYDR ARACHIDO CORTISON DEHYDROI INDOLEPR LACTATE NACETYLA PCRESOLS
LOC440131	<0.0001	0.0003	12	X.1ARACHI V83_A V84_A X.3DEHYDR ARACHIDO CORTISON DEHYDROI INDOLEPR LACTATE NACETYLA PCRESOLS THEOPHYL
NFATC3	<0.0001	0.0002	2	GLUTAROY LYSINE
PLA2G15	<0.0001	<0.0001	2	GLUTAROY LYSINE
TACO1	<0.0001	0.0002	3	ASPARTAT ASPARTYL HWESASXX
CALR	<0.0001	<0.0001	3	ERYTHRIT GLUTAROY GLYCEROP
DAND5	<0.0001	<0.0001	5	ARABITOL ERYTHRIT GLUTAROY GLYCEROP OLEAMIDE
DNASE2	<0.0001	<0.0001	3	ERYTHRIT GLUTAROY GLYCEROP
FARSA	<0.0001	<0.0001	5	ARABITOL ERYTHRIT GLUTAROY GLYCEROP OLEAMIDE
GADD45GIP1	<0.0001	<0.0001	5	ARABITOL ERYTHRIT FRUCTOSE GLUTAROY OLEAMIDE
GCDH	<0.0001	<0.0001	3	ERYTHRIT GLUTAROY GLYCEROP
KLF1	<0.0001	<0.0001	5	ARABITOL ERYTHRIT GLUTAROY GLYCEROP OLEAMIDE
MAST1	<0.0001	<0.0001	3	ERYTHRIT GLUTAROY GLYCEROP
NFIX	<0.0001	<0.0001	6	ARABITOL ERYTHRIT FRUCTOSE GLUTAROY GLYCEROP OLEAMIDE
PRDX2	<0.0001	<0.0001	3	ERYTHRIT GLUTAROY GLYCEROP
RAD23A	<0.0001	<0.0001	5	ARABITOL ERYTHRIT GLUTAROY GLYCEROP OLEAMIDE
RTBDN	<0.0001	<0.0001	3	ERYTHRIT GLUTAROY GLYCEROP
SYCE2	<0.0001	<0.0001	5	ARABITOL ERYTHRIT GLUTAROY GLYCEROP OLEAMIDE
NTN5	<0.0001	0.0003	2	ADPSGEGD V285_A

Table D.3.: **Significant enriched new phenotype sets without possibility of replication in TwinsUK.** All new identified phenotype sets that were signlificantly enriched in KORA F4 but could not be replicated in TwinsUK data as the phenotype sets could not be tested in TwinsUK data. num: number of all metabolites in the metabolite set.

gene	P (KORA F4)	num	elements of the phenotype set
LOC389033	<0.0001	2	ARABITOL GLUCOSE
SNORA62	<0.0001	2	ARABITOL PARAXANT
ATXN7	<0.0001	2	STACHYDR THREONAT
C3orf49	<0.0001	2	STACHYDR THREONAT
THOC7	<0.0001	2	STACHYDR THREONAT
BOP1	<0.0001	2	X.15ANHYD OLEAMIDE
GPR172A	<0.0001	2	X.15ANHYD OLEAMIDE
HSF1	<0.0001	2	X.15ANHYD OLEAMIDE
KAT5	<0.0001	2	V82_A STACHYDR
RELA	<0.0001	2	V82_A STACHYDR
NUP107	<0.0001	2	LINOLEAM LYSINE
SLC35E3	<0.0001	2	LINOLEAM LYSINE
GPR68	<0.0001	2	V144_A LINOLEAM
MIR1197	<0.0001	2	X.4ACETAM STACHYDR
MIR299	<0.0001	2	X.4ACETAM STACHYDR
MIR323	<0.0001	2	X.4ACETAM STACHYDR
MIR329-1	<0.0001	2	X.4ACETAM STACHYDR
MIR329-2	<0.0001	2	X.4ACETAM STACHYDR
MIR380	<0.0001	2	X.4ACETAM STACHYDR
MIR411	<0.0001	2	X.4ACETAM STACHYDR
MIR758	<0.0001	2	X.4ACETAM STACHYDR
SNORD114-8	<0.0001	2	PROPIONY STACHYDR
SNORD114-9	<0.0001	2	PROPIONY STACHYDR
USP50	<0.0001	2	X.3CARBOX LINOLEAM
RAB8B	<0.0001	2	X.4ACETAM ARABITOL
RPS27L	<0.0001	2	X.4ACETAM ARABITOL
OR4F15	<0.0001	2	X.5DODECE OLEAMIDE
OR4F6	<0.0001	2	X.5DODECE OLEAMIDE
SGK494	<0.0001	2	V80_A ARABITOL
SUPT6H	<0.0001	2	V80_A ARABITOL
MED24	<0.0001	2	V61_A STACHYDR
MSL1	<0.0001	2	V61_A STACHYDR
ATP5A1	<0.0001	2	NONADECA OLEAMIDE
MED25	<0.0001	2	V188_A STACHYDR
PTOV1	<0.0001	2	V188_A STACHYDR
TSKS	<0.0001	2	V188_A STACHYDR
EIF2S2	<0.0001	2	ARABITOL PROHYDRO
SERINC3	<0.0001	2	ADSGEGDF LINOLEAM
YWHAB	<0.0001	2	V70_A STACHYDR

D.2. Biological Interpretation of New Loci

In the following, all genetic loci that were identified with PSEA on metabolite data and were not found in a single phenotype GWAS on the same data are described and set into biological context.

Genes Identified on Biocrates Data

***IL3*:** Interpretation given in section 5.4.

***MFSD2A, MYCL1*:** At these two overlapping genes on chromosome one the same phenotype set consisting of five lyso-phosphatidylcholines of medium chain length (C16 - C20) was identified, significantly enriched and replicated in TwinsUK data. The gene product of *MFSD2A* (Major Facilitator Superfamily Domain Containing 2) is a transmembrane protein that belongs to a large family of presumptive carbohydrate transporters. Although database analysis of homologs suggests this function, there is no definitive proof (Esnault *et al.*, 2008). It is known that *MFSD2A* is down regulated in human lung primary tumors and lung cancer cell lines (Spinola *et al.*, 2010). Moreover, a study described it as receptor for a retrovirus-derived protein (syncytin-2) (Esnault *et al.*, 2008). *MYCL1* is known to be a member of the MYC oncogene family and to be involved in genesis of small cell lung cancer (Xiong *et al.*, 2011). Carbohydrates are connected with the fatty acid metabolism via the citrate cycle and with lyso-phosphatidylcholines. Therefore, the lyso-phosphatidylcholines might support the hypothesis of *MFSD2A* being a carbohydrate transporter.

***PDCD6IP*:** The gene *PDCD6IP* was identified with an enrichment of a new identified phenotype set consisting of six different carnitines (acylcarnitine, hydroxy- and dicarboxyacylcarnitines) and three lyso-phosphatidylcholines. The chain length of the selected acylcarnitines and hydroxyacylcarnitines ranges between 8 and 18 and the chain length of the lyso-phosphatidylcholines is 16 respectively 18. This gene is named programmed cell death 6 interacting protein or alternatively ALG-2 (apoptosis-linked-gene-2 product)-interacting protein X (ALIX). It is known to be involved in various cellular processes, for

example within the ESCRT (endosomal sorting complex required for transport) pathway in abscission of cytokinesis, actin-based cytoskeleton assembly and assembly of retroviruses (Shibata *et al.*, 2008; Zhou *et al.*, 2010). Another study found, that the coded protein ALIX controls the function of the lysobisphosphatidic acid (LBPA) in formation of multivesicular liposomes (Matsuo *et al.*, 2004; Dikic, 2004). The elements of the enriched phenotype set might be involved in this reaction.

***C12orf75*:** At gene *C12orf75* a set of seven different phosphatidylcholines and sphingomyelins along with five acyl-, hydroxyacyl- or dicarboxyacylcarnitines were identified and significantly enriched. No gene function is known for *C12orf75*. The elements of the enriched phenotype set might hint to an involvement of *C12orf75* in lipid metabolism.

***DKFZp686O1327*:** A set of eleven metabolites of eight metabolite classes, were enriched at an uncharacterized region spanning 550 kb on chromosome two. Due to the diversity of the metabolites in the enriched metabolite set and the unknown gene funtion no presumption about the gene metabolite network could be made.

Genes Identified on Metabolon Data

***LINGO2*:** Interpretation given in section 5.4.

cytochrome P450 family 4: Interpretation given in section 5.4.

***CEP68, SLC1A4*:** The new identified phenotype set at the gene *CEP68* is a subset of the phenotype set at *SLC1A4*. Approximately 63 % respectively 65 % of the physical length of these gene regions is overlapping. Both new sets included different amino acids (L-alanine, D-2-aminobutyrate, L-glutamine, L-threonine). The set of *SLC1A4* contained a steroid (7-alpha-hydroxy-3-oxo-4-cholestenoate) in addition to these amino acids. *SLC1A4* is a part of the solute carrier family 1, member 4 that is known to function as amino acid transporter. Arriza *et al.* (1993) showed that *SLC1A4* is a neutral amino acid transporter with a broad specificity. Amino acids like alanine and threonine are effective substrates of the gene product of *SLC1A4*. Therefore, one could state that some of the identified elements in the new sets are cargo of this gene product.
SLC1A4 was found to be associated with multiple system atrophy through the L-serine

metabolism (Soma *et al.*, 2008). Although, for linkage analysis *SLC1A4* was proposed as candidate gene for bipolar disorder and schizophrenia, association analysis could not confirm this (Skowronek *et al.*, 2006). For *CEP68* less is known. It was found to be a possible susceptible gene for aspirin intolerance in asthmatics in a Korean cohort (Kim *et al.*, 2010).

***DIRAS3*:** The gene *DIRAS3* (*ARHI*) was identified for three metabolites of the glycerolipid metabolism (1-oleoylglycerophosphoethanolamine, glycerol and D-glycerate) and 3-methyl-2-oxopentanoate, which is part of the valine, leucine and isoleucine metabolism. *DIRAS3* is a member of the Ras superfamily. Various publications have discussed the gene as tumor suppressor in different types of cancer (Huang *et al.*, 2009; Zou *et al.*, 2011; Badgwell *et al.*, 2012; Yu *et al.*, 1999). *DIRAS3* is assumed to inhibit cell growth and to play a role in the carcioneogenesis (Huang *et al.*, 2009). Three metabolites in the identified set are involved in the glycerolipid metabolism. The remaining metabolite 3-methyl-2-oxopentanoate might be connected with the other metabolites via the citrate cycle. Possibly, the effect of *DIRAS3* on cell growth changes the demand on carbon sources or energy which modifies glycolysis respectively glycerolipid metabolism.

***GCDH and others*:** Several overlapping genes on chromosome 19 were identified (*GCDH, CALR, DNASE2, MAST1, PRDX2, RTBDN, DAND5, FARSA, KLF1, NFIX, RAD23A, SYCE2, GADD45GIP1*). All enriched sets for these genes include erythritol, glutaroyl carnitine and one glycerolipid (sn-glycero-3-phosphocholine), other sets include two additional sugars (β-D-fructose, L-arabitol) and the fatty acid oleamide. The gene product of *GCDH* is a glutaryl-CoA dehydrogenase. Glutaryl-CoA is involved in tryptophan and lysine degradation and in the fatty acid metabolism. In several steps acetyl-CoA can be created from the product of glutaryl-CoA dehydrogenase. Carnitine and lipids of the enriched phenotype sets could reflect the involvement of glutaryl-CoA dehydrogenase in the fatty acids metabolism. The sugars might stem from the connection of acetyl-CoA with glycolysis. Moreover, *GCDH* was shown to be associated with mean corpuscular haemoglobin levels (Ganesh *et al.*, 2009).

***miR138-1*:** At microRNA 138-1 (*miR138-1*) spanning approximately 150 kb on chro-

mosome three a set with five metabolites of different glycerolipids, creatinine and the peptide g-γ-glutamyl was identified. The miR138 was detected to reduce the expression of histone H2AX and it induces chromosomal instability after DNA damage (Wang *et al.*, 2011). MicroRNAs are small RNAs that play an important role in gene regulatory processes. Apart from very few microRNAs that show a specific expression, most microRNAs regulate different genes with various biological functions (Bartel, 2009). The inclusion of metabolites that are part of different pathways in the enriched set, let presume a function of miR138-1 as a modulator of expression of multiple genes.

***UBL3, LOC440131*:** The ubiquitin-like 3 (*UBL3*) gene was identified for a phenotype set including fatty acids, glycerolipids, amino acids, carnitine, lactate, two steroids and theophylline. The gene function is not known, but it was reported to be weakly associated with Biliary atresia (Garcia-Barceló *et al.*, 2010). Ubiquitin-like proteins are proteins that can be attached to various proteins modifying subsequent processes (Kerscher *et al.*, 2006). This kind of protein-based modification was first described for ubiquitin. Numerous reactions can be influenced by one ubiquitin-like protein (Hochstrasser, 2009). This could explain the enriched phenotype set, which includes different types of metabolites.

Bibliography

Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10: 47.

Andrew, T., Hart, D.J., Snieder, H., de Lange, M., Spector, T.D. and MacGregor, A.J. (2001). Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. *Twin Res*, 4(6): 464–477.

Arriza, J.L., Kavanaugh, M.P., Fairman, W.A., Wu, Y.N., Murdoch, G.H., North, R.A. and Amara, S.G. (1993). Cloning and expression of a human neutral amino acid transporter with structural similarity to the glutamate transporter gene family. *J Biol Chem*, 268(21): 15329–15332.

Badgwell, D.B., Lu, Z., Le, K., Gao, F., Yang, M., Suh, G.K., Bao, J.J., Das, P., Andreeff, M., Chen, W., Yu, Y., Ahmed, A.A., Liao, W.S.L. and Bast, R.C. (2012). The tumor-suppressor gene ARHI (DIRAS3) suppresses ovarian cancer cell migration through inhibition of the Stat3 and FAK/Rho signaling pathways. *Oncogene*, 31(1): 68–79.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2): 215–233.

Bauer, D.E., Hatzivassiliou, G., Zhao, F., Andreadis, C. and Thompson, C.B. (2005). ATP citrate lyase is an important component of cell growth and transformation. *Oncogene*, 24(41): 6314–6322.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc*, 57(1): 298–300.

Benyamin, B., McRae, A.F., Zhu, G., Gordon, S., Henders, A.K., Palotie, A., Peltonen, L., Martin, N.G., Montgomery, G.W., Whitfield, J.B. and Visscher, P.M. (2009). Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. *Am J Hum Genet*, 84(1): 60–65.

Borry, P., Cornel, M.C. and Howard, H.C. (2010). Where are you going, where have you been: a recent history of the direct-to-consumer genetic testing market. *J Community Genet*, 1(3): 101–106.

Chambers, J.C., Zhang, W., Li, Y., Sehmi, J., Wass, M.N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M.I., Peltonen, L., Freimer, N.B., Srai, S.K., Maxwell, P.H., Sternberg, M.J.E., Ruokonen, A., Abecasis, G., Jarvelin, M.R., Scott, J., Elliott, P. and Kooner, J.S. (2009). Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nat Genet*, 41(11): 1170–1172.

Chaudhary, K.R., Batchu, S.N. and Seubert, J.M. (2009). Cytochrome P450 enzymes and the heart. *IUBMB Life*, 61(10): 954–960.

Deberardinis, R.J., Lum, J.J. and Thompson, C.B. (2006). Phosphatidylinositol 3-kinase-dependent modulation of carnitine palmitoyltransferase 1A expression regulates lipid metabolism during hematopoietic cell growth. *J Biol Chem*, 281(49): 37372–37380.

Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M. and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 26(9): 1205–1210.

Dikic, I. (2004). ALIX-ing phospholipids with endosome biogenesis. *Bioessays*, 26(6): 604–607.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann Appl Stat*, 1(1): 107–129.

Esnault, C., Priet, S., Ribet, D., Vernochet, C., Bruls, T., Lavialle, C., Weissenbach, J. and Heidmann, T. (2008). A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2. *Proc Natl Acad Sci U S A*, 105(45): 17532–17537.

Evans, A.M., DeHaven, C.D., Barrett, T., Mitchell, M. and Milgram, E. (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem*, 81(16): 6656–6667.

Ferreira, M.A.R. and Purcell, S.M. (2009). A multivariate test of association. *Bioinformatics*, 25: 132–133.

Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., Anderson, C.A., Bis, J.C., Bumpstead, S., Ellinghaus, D., Festen, E.M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C.G., Montgomery, G.W., Prescott, N.J., Raychaudhuri, S., Rotter, J.I. *et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*, 42(12): 1118–1125.

Gainer, J.V., Bellamine, A., Dawson, E.P., Womble, K.E., Grant, S.W., Wang, Y., Cupples, L.A., Guo, C.Y., Demissie, S., O'Donnell, C.J., Brown, N.J., Waterman, M.R. and Capdevila, J.H. (2005). Functional variant of CYP4A11 20-hydroxyeicosatetraenoic acid synthase is associated with essential hypertension. *Circulation*, 111(1): 63–69.

Ganesh, S.K., Zakai, N.A., van Rooij, F.J.A., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.H., Kottgen, A., Glazer, N.L., Dehghan, A., Kuhnel, B., Aspelund, T., Yang,

Q., Tanaka, T., Jaffe, A., Bis, J.C.M., Verwoert, G.C., Teumer, A., Fox, C.S., Guralnik, J.M., Ehret, G.B., Rice, K., Felix, J.F., Rendon, A., Eiriksdottir, G. *et al.* (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE consortium. *Nat Genet*, 41(11): 1191–1198.

Garcia-Barceló, M.M., Yeung, M.Y., Miao, X.P., Tang, C.S.M., Cheng, G., Chen, G., So, M.T., Ngan, E.S.W., Lui, V.C.H., Chen, Y., Liu, X.L., Hui, K.J.W.S., Li, L., Guo, W.H., Sun, X.B., Tou, J.F., Chan, K.W., Wu, X.Z., Song, Y.Q., Chan, D., Cheung, K., Chung, P.H.Y., Wong, K.K.Y., Sham, P.C., Cherny, S.S. *et al.* (2010). Genome-wide association study identifies a susceptibility locus for biliary atresia on 10q24.2. *Hum Mol Genet*, 19(14): 2917–2925.

Gieger, C., Geistlinger, L., Altmaier, E., de Angelis, M.H., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J., Illig, T. and Suhre, K. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, 4(11): e1000282.

Guo, Y.F., Li, J., Chen, Y., Zhang, L.S. and Deng, H.W. (2009). A new permutation strategy of pathway-based approach for genome-wide association study. *BMC Bioinformatics*, 10: 429.

Gupta, M., Cheung, C.L., Hsu, Y.H., Demissie, S., Cupples, L.A., Kiel, D.P. and Karasik, D. (2011). Identification of homogeneous genetic architecture of multiple genetically correlated traits by block clustering of genome-wide associations. *J Bone Miner Res*, 26(6): 1261–1271.

He, X. and Zhang, J. (2006). Toward a molecular understanding of pleiotropy. *Genetics*, 173(4): 1885–1891.

Hentze, M.W., Muckenthaler, M.U., Galy, B. and Camaschella, C. (2010). Two to tango: regulation of mammalian iron metabolism. *Cell*, 142(1): 24–38.

Hochstrasser, M. (2009). Origin and function of ubiquitin-like proteins. *Nature*, 458(7237): 422–429.

Huang, J., Johnson, A.D. and O'Donnell, C.J. (2011). PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics*, 27(9): 1201–1206.

Huang, J., Lin, Y., Li, L., Qing, D., Teng, X.M., Zhang, Y.L., Hu, X., Hu, Y., Yang, P. and guang Han, Z. (2009). ARHI, as a novel suppressor of cell growth and downregulated in human hepatocellular carcinoma, could contribute to hepatocarcinogenesis. *Mol Carcinog*, 48(2): 130–140.

Huang, Y., Xu, H., Calian, V. and Hsu, J.C. (2006). To permute or not to permute. *Bioinformatics*, 22(18): 2244–2248.

Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B.S., Mewes, H.W., Meitinger, T., de Angelis, M.H., Kronenberg, F., Soranzo, N., Wichmann, H.E., Spector, T.D., Adamski, J. and Suhre, K. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*, 42(2): 137–141.

Kaizar, E.E., Li, Y. and Hsu, J.C. (2011). Permutation multiple tests of binary features do not uniformly control error rates. *Journal of the American Statistical Association*, 106(495): 1067–1074.

Kerscher, O., Felberbaum, R. and Hochstrasser, M. (2006). Modification of proteins by ubiquitin and ubiquitin-like proteins. *Annu Rev Cell Dev Biol*, 22: 159–180.

Kilpeläinen, T.O., Zillikens, M.C., Stančákova, A., Finucane, F.M., Ried, J.S., Langenberg, C., Zhang, W., Beckmann, J.S., Luan, J., Vandenput, L., Styrkarsdottir, U., Zhou, Y., Smith, A.V., Zhao, J.H., Amin, N., Vedantam, S., Shin, S.Y., Haritunians, T., Fu, M., Feitosa, M.F., Kumari, M., Halldorsson, B.V., Tikkanen, E., Mangino, M., Hayward, C. *et al.* (2011). Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nat Genet*, 43(8): 753–760.

Kim, J.H., Park, B.L., Cheong, H.S., Bae, J.S., Park, J.S., Jang, A.S., Uh, S.T., Choi, J.S., Kim, Y.H., Kim, M.K., Choi, I.S., Cho, S.H., Choi, B.W., Park, C.S. and Shin,

H.D. (2010). Genome-wide and follow-up studies identify CEP68 gene variants associated with risk of aspirin-intolerant asthma. *PLoS One*, 5(11): e13818.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J. and Theis, F.J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5: 21.

Kuhajda, F.P., Jenner, K., Wood, F.D., Hennigar, R.A., Jacobs, L.B., Dick, J.D. and Pasternack, G.R. (1994). Fatty acid synthesis: a potential selective target for antineoplastic therapy. *Proc Natl Acad Sci U S A*, 91(14): 6379–6383.

Kullo, I.J., Ding, K., Jouni, H., Smith, C.Y. and Chute, C.G. (2010). A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS One*, 5(9): e13011.

Liu, J., Pei, Y., Papasian, C.J. and Deng, H.W. (2009). Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet Epidemiol*, 33(3): 217–227.

Lopez, A.F., Dyson, P.G., To, L.B., Elliott, M.J., Milton, S.E., Russell, J.A., Juttner, C.A., Yang, Y.C., Clark, S.C. and Vadas, M.A. (1988). Recombinant human interleukin-3 stimulation of hematopoiesis in humans: loss of responsiveness with differentiation in the neutrophilic myeloid series. *Blood*, 72(5): 1797–1804.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7): 499–511.

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7): 906–913.

Matsuo, H., Chevallier, J., Mayran, N., Blanc, I.L., Ferguson, C., Fauré, J., Blanc, N.S., Matile, S., Dubochet, J., Sadoul, R., Parton, R.G., Vilbois, F. and Gruenberg, J.

(2004). Role of LBPA and Alix in multivesicular liposome formation and endosome organization. *Science*, 303(5657): 531–534.

Meyer, C.G., Fernandes, M.H.C., Intemann, C.D., Kreuels, B., Kobbe, R., Kreuzberg, C., Ayim, M., Ruether, A., Loag, W., Ehmen, C., Adjei, S., Adjei, O., Horstmann, R.D. and May, J. (2011). IL3 variant on chromosomal region 5q31-33 and protection from recurrent malaria attacks. *Hum Mol Genet*, 20(6): 1173–1181.

Nichols, M.J., Saavedra-Matiz, C.A., Pass, K.A. and Caggana, M. (2008). Novel mutations causing medium chain acyl-CoA dehydrogenase deficiency: under-representation of the common c.985 A $>$ G mutation in the New York state population. *Am J Med Genet A*, 146A(5): 610–619.

O'Brien, P. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40: 1079–1087.

Oexle, K., Ried, J.S., Hicks, A.A., Tanaka, T., Hayward, C., Bruegel, M., Gögele, M., Lichtner, P., Müller-Myhsok, B., Döring, A., Illig, T., Schwienbacher, C., Minelli, C., Pichler, I., Fiedler, G.M., Thiery, J., Rudan, I., Wright, A.F., Campbell, H., Ferrucci, L., Bandinelli, S., Pramstaller, P.P., Wichmann, H.E., Gieger, C., Winkelmann, J. *et al.* (2011). Novel association to the proprotein convertase PCSK7 gene locus revealed by analysing soluble transferrin receptor (stfr) levels. *Hum Mol Genet*, 20(5): 1042–1047.

Ohta, T., Masutomi, N., Tsutsui, N., Sakairi, T., Mitchell, M., Milburn, M.V., Ryals, J.A., Beebe, K.D. and Guo, L. (2009). Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. *Toxicol Pathol*, 37(4): 521–535.

Orkin, S.H. and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4): 631–644.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. and Sham, P.C. (2007). PLINK: a tool

set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3): 559–575.

Ried, J.S., Döring, A., Oexle, K., Meisinger, C., Winkelmann, J., Klopp, N., Meitinger, T., Peters, A., Suhre, K., Wichmann, H.E. and Gieger, C. (2012). PSEA: phenotype set enrichment analysis - a new method for analysis of multiple phenotypes. *Genet Epidemiol*, 36(3): 244–252.

Segrè, A.V., D.I.A.G.R.A.M. Consortium, M.A.G.I.C. investigators, Groop, L., Mootha, V.K., Daly, M.J. and Altshuler, D. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet*, 6(8): e1001058.

Shibata, H., Suzuki, H., Kakiuchi, T., Inuzuka, T., Yoshida, H., Mizuno, T. and Maki, M. (2008). Identification of Alix-type and Non-Alix-type ALG-2-binding sites in human phospholipid scramblase 3: differential binding to an alternatively spliced isoform and amino acid-substituted mutants. *J Biol Chem*, 283(15): 9623–9632.

Shriner, D. (2012). Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Front Genet*, 3: 1.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J.G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J.F. and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet*, 89(5): 607–618.

Skowronek, M.H., Georgi, A., Jamra, R.A., Schumacher, J., Becker, T., Schmael, C., Paul, T., Deschner, M., Höfels, S., Wulff, M., Schwarz, M., Klopp, N., Illig, T., Propping, P., Cichon, S., Nöthen, M.M., Schulze, T.G. and Rietschel, M. (2006). No association between genetic variants at the ASCT1 gene and schizophrenia or bipolar disorder in a german sample. *Psychiatr Genet*, 16(6): 233–234.

Soma, H., Yabe, I., Takei, A., Fujiki, N., Yanagihara, T. and Sasaki, H. (2008). Associations between multiple system atrophy and polymorphisms of SLC1A4, SQSTM1, and EIF4EBP1 genes. *Mov Disord*, 23(8): 1161–1167.

Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., Salo, P., Voight, B.F., Burns, P., Laskowski, R.A., Xue, Y., Menzel, S., Altshuler, D., Bradley, J.R., Bumpstead, S., Burnett, M.S., Devaney, J., Döring, A., Elosua, R., Epstein, S.E., Erber, W. *et al.* (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet*, 41(11): 1182–1190.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J., Mägi, R., Randall, J.C., Vedantam, S., Winkler, T.W., Qi, L., Workalemahu, T., Heid, I.M., Steinthorsdottir, V., Stringham, H.M., Weedon, M.N., Wheeler, E., Wood, A.R., Ferreira, T., Weyant, R.J., Segrè, A.V., Estrada, K. *et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11): 937–948.

Spinola, M., Falvella, F.S., Colombo, F., Sullivan, J.P., Shames, D.S., Girard, L., Spessotto, P., Minna, J.D. and Dragani, T.A. (2010). MFSD2A is a novel lung tumor suppressor gene modulating cell cycle and matrix attachment. *Mol Cancer*, 9: 62.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43): 15545–15550.

Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., C. A. R. D. IoG. R. A. M., Deloukas, P., Erdmann, J., Grundberg, E., Hammond, C.J., de Angelis, M.H., Kastenmüller, G., Köttgen, A., Kronenberg, F., Mangino, M., Meisinger, C., Meitinger, T., Mewes, H.W., Milburn, M.V., Prehn, C., Raffler, J., Ried, J.S., Römisch-Margl, W. *et al.* (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362): 54–60.

Tanaka, T., Roy, C.N., Yao, W., Matteini, A., Semba, R.D., Arking, D., Walston, J.D., Fried, L.P., Singleton, A., Guralnik, J., Abecasis, G.R., Bandinelli, S., Longo, D.L. and

Ferrucci, L. (2010). A genome-wide association analysis of serum iron concentrations. *Blood*, 115(1): 94–96.

Tintle, N.L., Borchers, B., Brown, M. and Bekmetjev, A. (2009). Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. *BMC Proc*, 3 Suppl 7: S96.

Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M. and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, 4(10): e1000214.

Wade, C.H. and Wilfond, B.S. (2006). Ethical and clinical practice considerations for genetic counselors related to direct-to-consumer marketing of genetic tests. *Am J Med Genet C Semin Med Genet*, 142C(4): 284–92, discussion 293.

Wang, K., Li, M. and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 81(6): 1278–1283.

Wang, K., Li, M. and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*, 11(12): 843–854.

Wang, Y., Huang, J.W., Li, M., Cavenee, W.K., Mitchell, P.S., Zhou, X., Tewari, M., Furnari, F.B. and Taniguchi, T. (2011). MicroRNA-138 modulates DNA damage response by repressing histone H2AX expression. *Mol Cancer Res*, 9(8): 1100–1111.

Wichmann, H.E., Gieger, C. and Illig, T. (2005). KORA-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 67: S26–S30.

Wu, Y.W., Prakash, K.M., Rong, T.Y., Li, H.H., Xiao, Q., Tan, L.C., Au, W.L., qing Ding, J., di Chen, S. and Tan, E.K. (2011). Lingo2 variants associated with essential tremor and Parkinson's disease. *Hum Genet*, 129(6): 611–615.

Xiong, F., Wu, C., Chang, J., Yu, D., Xu, B., Yuan, P., Zhai, K., Xu, J., Tan, W. and

Lin, D. (2011). Genetic variation in an miRNA-1827 binding site in MYCL1 alters susceptibility to small-cell lung cancer. *Cancer Res*, 71(15): 5175–5181.

Yang, Q., Wu, H., Guo, C.Y. and Fox, C.S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet Epidemiol*, 34(5): 444–454.

Yu, Y., Xu, F., Peng, H., Fang, X., Zhao, S., Li, Y., Cuevas, B., Kuo, W.L., Gray, J.W., Siciliano, M., Mills, G.B. and Bast, R.C. (1999). NOEY2 (ARHI), an imprinted putative tumor suppressor gene in ovarian and breast carcinomas. *Proc Natl Acad Sci U S A*, 96(1): 214–219.

Zhou, X., Si, J., Corvera, J., Gallick, G.E. and Kuang, J. (2010). Decoding the intrinsic mechanism that prohibits ALIX interaction with ESCRT and viral proteins. *Biochem J*, 432(3): 525–534.

Zou, C.F., Jia, L., Jin, H., Yao, M., Zhao, N., Huan, J., Lu, Z., Bast, R.C., Feng, Y. and Yu, Y. (2011). Re-expression of ARHI (DIRAS3) induces autophagy in breast cancer cells and enhances the inhibitory effect of paclitaxel. *BMC Cancer*, 11: 22.

Acknowledgement

Study Acknowledgements

The KORA Augsburg studies were financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany and supported by grants from the German Federal Ministry of Education and Research (BMBF). Part of this work was financed by the German National Genome Research Network (NGFN). Our research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ.

The TwinsUK study was funded by the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F2-2008-201865-GEFOS and (FP7/2007-2013), ENGAGE project grant agreement HEALTH-F4-2007-201413 and the FP-5 GenomEUtwin Project (QLG2-CT-2002-01254). The study also receives support from the Dept of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London. TDS is an NIHR senior Investigator. The project also received support from a Biotechnology and Biological Sciences Research Council (BBSRC) project grant (G20234). The authors acknowledge the funding and support of the National Eye Institute via an NIH/CIDR genotyping project (PI: Terri Young). We thank the staff from the Genotyping Facilities at the Wellcome Trust Sanger Institute for sample preparation, Quality Control and Genotyping led by Leena Peltonen and Panos Deloukas; Le Centre National de Génotypage, France, led by Mark Lathrop, for genotyping; Duke University, North Carolina, USA, led by David Goldstein, for genotyping; and the Finnish Institute of Molecular Medicine, Finnish Genome

Center, University of Helsinki, led by Aarno Palotie. Genotyping was also performed by CIDR as part of an NEI/NIH project grant.

DEISA Acknowledgement

I thank the DEISA Consortium (`www.deisa.eu`), co-funded through the EU FP6 project RI-031513 and the FP7 project RI-222919, for support within the DEISA Extreme Computing Initiative. Moreover, we thank the DEISA-group of LRZ for technically support, especially Dr. Kamen Beronov, Siew Hoon Leong and Markus Michael Müller.

Personal Acknowledgement

I thank my supervisor Prof. Dr. Dr. Wichmann for giving me the opportunity to conduct this PhD-project. I thank my co-supervisor Dr. Christian Gieger for his continuous support during all stages of the project. I wish to thank Prof. Dr. Karsten Suhre for his very helpful ideas and for giving me the opportunity to participate in the DEISA project for high performance computing. For providing data from the TwinsUK study I thank PhD So-Youn Shin and Dr. Nicole Soranzo from the Wellcome Trust Sanger Institute in Hinxton, UK. Their data enabled the important replication of my results. I thank Jan Krumsiek from the Helmholtz Zentrum Munich, who helped with the calculation of metabolite sets with his Gaussian graphical modelling method. I whish to thank Prof. Dr. Strauch, head of the institute of Genetic Epidemiology at the Helmholtz Zentrum Munich and my colleagues for the good working atmosphere, helpful advices and best collaboration.

Last but most importantly, I wish to thank my wonderful family for accompanying me through all ups and downs of this PhD-project. I thank my parents, who always support me in all my pursuits and are on my side with encouraging words and deeds.[1] I thank my husband for his support, his ideas, his patience, his understanding and his love.

[1] Zuletzt aber am wichtigsten, möchte ich meiner wundervollen Familie danken, dass sie mich durch alle Höhen und Tiefen der Promotion begleitet hat. Ich danke meinen Eltern, die mich immer in jeder Hinsicht unterstützen und mir mit Rat und Tat zur Seite stehen.