From
the Institute of Genetic Epidemiology, Helmholtz Zentrum München
Director: Prof. Dr. Konstantin Strauch
in cooperation with
the Institute of Medical Informatics, Biometry and Epidemiology,
Ludwig-Maximilians-University of Munich
Director: Prof. Dr. Ulrich Mansmann
Chair of Epidemiology: Prof. Dr. Dr. H.-Erich Wichmann (emeritus)

Statistical incorporation of metabolites in the genome-wide association study approach

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

by
Ann-Kristin Petersen

from
Flensburg

2012

With approval of the Faculty of Medicine
of the Ludwig-Maximilians-University of Munich

Supervisor:	Prof. Dr. Dr. H.-Erich Wichmann
Co-referee:	Prof. Dr. Ralf Zimmer
Dean:	Prof. Dr. med. Dr. h. c. Maximilian Reiser, FACR, FRCR
Date of oral examination:	02.10.2013

To my parents

This thesis is based on results of the following publications:

Petersen AK, Stark K, Musameh MD, Nelson CP, Römisch-Margl W, Kremer W, Raffler J, Krug S, Skurk T, Rist MJ, Daniel H, Hauner H, Adamski J, Tomaszewski M, Döring A, Peters A, Wichmann HE, Kaess BM, Kalbitzer HR, Huber F, Pfahlert V, Samani NJ, Kronenberg F, Dieplinger H, Illig T, Hengstenberg C, Suhre K, Gieger C, Kastenmüller G. **Genetic associations with lipoprotein subfractions provide information on their biological nature.** *Hum Mol Genet.* 2012; 21(6):1433-1443.
Permission for reuse of this publication in this thesis was granted by Oxford University Press.

Suhre K, Shin SY, **Petersen AK**, Mohney RP, Meredith D, Wägele B, Altmaier E, CARDIoGRAM, Deloukas P, Erdmann J, Grundberg E, Hammond CJ, de Angelis MH, Kastenmüller G, Köttgen A, Kronenberg F, Mangino M, Meisinger C, Meitinger T, Mewes HW, Milburn MV, Prehn C, Raffler J, Ried JS, Römisch-Margl W, Samani NJ, Small KS, Wichmann HE, Zhai G, Illig T, Spector TD, Adamski J, Soranzo N, Gieger C. **Human metabolic individuality in biomedical and pharmaceutical research.** *Nature.* 2011; 477(7362):54-60.
Permission for reuse of this publication in this thesis was granted by Nature Publishing Group.

Petersen AK, Krumsiek J, Wägele B, Theis FJ, Wichmann HE, Gieger C, Suhre K. **On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies.** *BMC Bioinformatics.* 2012; 13:120.
Permission for reuse of this publication in this thesis is according to the Creative Commons Attribution License.

Contents

Abbreviations

AU	approximately unbiased
BMI	body mass index
CHOD-PAP	cholesterol oxidase-p-aminophenazone
EDTA	ethylene diaminetetraacetic acid
eGFR	estimated glomerular filtration rate
GC	gas chromatography
GGMs	Gaussian graphical models
GPO-PAP	glycerol phosphate oxidase-p-aminophenazone
GRAPHIC	Genetic Regulation of Arterial Pressure of Humans in the Community
GWAS	genome-wide association studies
HDL	high density lipoprotein
HDL-C	high density lipoprotein cholesterol
HMDB	Human Metabolome Database
HuMet	Human Metabolome
IDL	intermediate density lipoprotein
KEGG	Kyoto Encyclopedia of Genes and Genomes
KORA	Cooperative Health Research in the Region of Augsburg

LC	liquid chromatography
LD	linkage disequilibrium
LDL	low density lipoprotein
LDL-C	low density lipoprotein cholesterol
MAF	minor allele frequency
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NMR	nuclear magnetic resonance
SMPDB	Small Molecule Pathway Database
SNP	single nucleotide polymorphism
TC	total cholesterol
TG	triglycerides
VLDL	very low density lipoprotein

List of Figures

List of Tables

Summary (English)

Epidemiological studies investigate complex diseases of which most have a predisposition through genetical factors, for example type 2 diabetes or cardiovascular diseases. In order to discover genes involved in the disease aetiology, genome-wide association studies (GWAS) are the state-of-the-art method. Hitherto, GWAS comprise of up to 250 000 samples but despite these large sample sizes only a fraction of the estimated heritability of the analysed phenotypes can be explained by the discovered genes, so far. Moreover, the genes detected in GWAS have to be further investigated to better understand biochemical processes underlying the association. A promising instrument to gain further insight is the analysis of metabolites. Metabolomics is the evolving field of measuring endogenous organic compounds of a cell or body fluid. As metabolites are downstream products of genetic processes, they are considered to exceed other phenotypes in power. Recently, some GWAS with metabolomics have been conducted and revealed promising results analysing ratios between metabolite concentrations (metabolite ratios). To decide whether a metabolite ratio carries more information than the two corresponding single metabolite concentrations alone, the p-gain was introduced as an objective measure. The p-gain is defined as the quotient of the smallest of the association p-values of the single metabolite concentrations to the association p-value of the metabolite ratio.

In this thesis, two procedures for the incorporation of metabolites in the GWAS approach are presented and applied to different metabolomics data sets. In addition, a statistical exploration of the p-gain is carried out to improve the examination of metabolite ratios. In the first of the two presented procedures, metabolites are used for an in-depth analysis of genetic candidate loci which have already been discovered in GWAS of clinically relevant phenotypes. In the second procedure, metabolites are used to discover new genetic loci through conduction of metabolomics GWAS. In a follow-up analysis, these novel loci should be further analysed together with clinically relevant phenotypes. As application of the first procedure, we conducted an analysis of 95 known serum lipid loci using 15 lipopro-

tein subfractions. We revealed significant associations for eight of the loci and thus gained further insight into different lipid pathways. As an application of the second procedure, we conducted GWAS of more than 250 metabolites as well as all pair-wise ratios, in total over 37 000 metabolic traits. These analyses revealed 37 loci which lead to further insight into various pathways of the human metabolism. In a follow-up analysis, some loci also showed associations with clinically relevant phenotypes. Finally, we determined the distribution of the p-gain and derived critical values through extensive statistical exploration. In conjunction with this, we demonstrated the power of the p-gain approach through a pathway enrichment analysis.

In conclusion, this thesis shows by concrete examples that both procedures for the incorporation of metabolomics data in the GWAS approach confirm and extend current knowledge about genetics underlying various biochemical pathways as well as discusses the advantages and limitations of both procedures and improves the examination of metabolite ratios.

Summary (German)

In epidemiologischen Studien werden komplexe Erkrankungen untersucht, von denen viele eine Prädisposition bezüglich genetischer Komponenten haben, z.B. Typ 2 Diabetes oder kardiovaskuläre Erkrankungen. Die Standardmethode für die Identifizierung von Genen, die in der Ätiologie von Krankheiten eine wichtige Rolle spielen, sind Genom-weite Assoziationsstudien (GWAS). In den zurzeit größten GWAS werden Daten von bis zu 250 000 Individuen ausgewertet. Trotz dieser großen Stichprobenumfänge wird bisher nur ein kleiner Teil der Erblichkeit von Phänotypen durch die entdeckten Gene erklärt. Neben der Identifizierung der Gene, sind die biochemischen Zusammenhänge zwischen den Genen und der Krankheit aufzuklären. Ein vielversprechender Weg hierfür ist die Analyse von Metaboliten. Metabolomics ist ein sich entwickelndes Gebiet, in dem endogene organische Komponenten einer Zelle oder Körperflüssigkeit gemessen werden. Da die Metabolite Produkte von genetischen Prozessen sind, birgt die Analyse von Metabolitendaten eine höhere Power als von anderen Phänotypen. Bisherige GWAS mit Metabolitendaten führten bereits zu sehr vielversprechenden Ergebnissen in der Analyse von Quotienten von Metabolitenkonzentrationen (Metabolitenquotienten). Um zu bestimmen, ob ein Metabolitenquotient mehr Informationen enthält als die beiden zugehörigen Metabolitenkonzentrationen alleine wurde der p-gain als objektives Maß eingeführt. Der p-gain ist definiert als Quotient des kleinsten p-Wertes der Assoziationen der Metabolitenkonzentrationen zum p-Wert der Assoziation des Metabolitenquotienten.

In dieser Dissertation werden zwei Verfahren zur Einbettung von Metaboliten in den GWAS Ansatz vorgestellt und auf verschiedene Datensätze angewendet. Darüber hinaus wird eine statistische Analyse des p-gains durchgeführt, um die Auswertung von Metabolitenquotienten zu verbessern. Die Idee des ersten der beiden vorgestellten Verfahren ist es, die Metaboliten für eine weiterführende Analyse von bereits bekannten genetischen Loci zu verwenden. Im Gegensatz dazu werden in dem zweiten vorgestellten Verfahren neue genetische Loci in GWAS mit Metabolitendaten entdeckt. In Folgeanalysen werden diese neuen Loci als Kan-

didatenloci bei Analysen mit klinisch relevanten Phänotypen weiter ausgewertet. Als Anwendung des ersten Verfahrens haben wir 95 bekannte Lipidloci mit Hilfe von 15 Lipoproteinsubklassen näher untersucht. Diese Analyse brachte für acht Loci einen tieferen Einblick in Zusammenhänge verschiedener Lipidstoffwechselwege. Als Anwendung des zweiten Verfahrens haben wir mehr als 250 Metabolite, sowie alle paarweisen Metabolitenquotienten analysiert, insgesamt mehr als 37 000 Metabolitenphänotypen. Diese Analyse hat 37 assoziierte Loci hervorgebracht, die neue Einblicke in verschiedene Stoffwechselwege geliefert haben. Darüber hinaus konnten für einige dieser Loci zusätzliche Assoziationen mit klinisch relevanten Phänotypen gezeigt werden. Abschließend haben wir für die statistische Auswertung des p-gains dessen Verteilung bestimmt, sowie zugehörige kritische Werte hergeleitet. Um die Relevanz des p-gain Konzeptes zu zeigen wurde außerdem nachgewiesen, dass für Metabolitenquotienten mit signifikantem p-gain die zugehörigen einzelnen Metabolitenkonzentrationen vermehrt zu einem gemeinsamen Stoffwechselweg gehören.

Insgesamt zeigt diese Dissertation an konkreten Beispielen, dass beide vorgestellte Verfahren zur Einbeziehung von Metaboliten in den GWAS Ansatz aktuelles Wissen über genetische und biochemische Prozesse verschiedener Stoffwechselwege sowohl bestätigen als auch erweitern. Darüber hinaus werden in dieser Dissertation die Vor- und Nachteile der beiden Verfahren diskutiert und die Auswertung von Metabolitenquotienten verbessert.

1. Introduction

Complex diseases such as type 2 diabetes or cardiovascular diseases are an increasing global health burden. According to the World Health Organisation (2011a,b), 346 million people worldwide suffer from type 2 diabetes whereas cardiovascular diseases are the number one cause of death globally. Elucidation of the aetiology of complex diseases in conjunction with an improvement of preventive medicine is an aim of epidemiological studies. In these studies, the disease itself as well as related risk factors are investigated. Furthermore, because a genetical predisposition exists for most complex diseases, the identification of genes involved in the disease aetiology is essential. For this purpose, genome-wide association studies (GWAS) are the state-of-the-art method. In order to gain further insight into genetical and biochemical mechanisms underlying a disease, this thesis expands the GWAS approach by incorporating metabolites as intermediate phenotypes between the genes and diseases.

1.1 Genome-wide association studies

GWAS is the hypothesis-free approach of statistically testing associations between a phenotype and millions of genetic variants, predominantly single nucleotide polymorphisms (SNPs). The underlying idea of GWAS is that a number of common SNPs are causal for a complex disease. Therefore, it is expected that differences in frequency for these SNPs can be detected between cases and controls (McCarthy *et al.*, 2008; Pearson and Manolio, 2008). The first GWAS were conducted in 2007 for diseases such as type 2 diabetes, Crohn's disease, Prostate cancer or coronary artery disease (Sladek *et al.*, 2007; Libioulle *et al.*, 2007; Yeager *et al.*, 2007; Burton *et al.*, 2007). These GWAS comprised of 500 to 2000 cases and 600 to 3000 controls and revealed up to nine associated genomic regions. In the meantime, GWAS were also conducted for many quantitative traits which are risk factors for various diseases. So far, a total of 1449 GWAS for 237 different traits are published (Hindorff *et al.*, 2011). The significant results

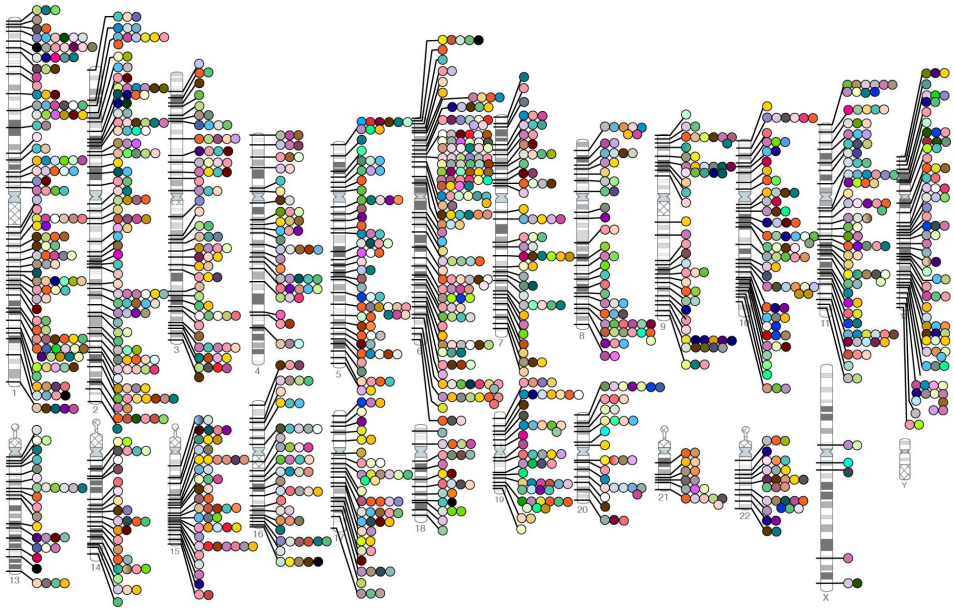

Figure 1.1: Published GWAS by June 2011. This Figure depicts significant associations (p-value $< 5 \times 10^{-8}$) detected in 1449 GWAS on 237 traits together with their location on the human genome. The 237 traits are colour coded. Courtesy: National Human Genome Research Institute (Hindorff *et al.*, 2011).

of these GWAS and their location on the genome are displayed in Figure 1.1. Although GWAS are a very popular method to reveal novel risk loci, one drawback is the small effect size of SNPs. This results mostly in an explained variance of less than 1 % or an odds ratio smaller than 1.2 (De Bakker *et al.*, 2008). Therefore, large sample sizes are needed to detect significant associations. Enlarging the sample size is often achieved through conduction of meta-analyses where multiple teams carry out the same analysis in different cohorts and combine the results afterwards (Zeggini and Ioannidis, 2009; Thompson *et al.*, 2011). Currently, the largest meta-analyses comprise of up to 250 000 samples, e.g. for body mass index (BMI), height or serum lipids (Speliotes *et al.*, 2010; Lango Allen *et al.*, 2010; Teslovich *et al.*, 2010). In these GWAS, 18 loci associated with BMI, 180 with height and 95 with serum lipids were detected. Together, these loci explain 3 % of the genetic variance of BMI, 13 % of height and 25 % – 30 % of serum lipids. These numbers show that a noteworthy proportion of the estimated heritability of these traits remains unexplained. This problem of the missing heritability is a widely discussed topic. Among the suspected reasons are undetected rare mutations which are not tagged well by common SNPs, common variants with a low penetrance, other genomic variations such as copy number variants, gene-gene and gene-environment interactions as well as inaccurate heritability estimates

(Maher, 2008). Analyses to discover some part of the missing heritability address the effects of many SNPs simultaneously. For example, analysing 294 831 SNPs together in one regression model can explain 45 % of the genetic variance of height. Additionally, some more of the unexplained heritability might be explained by incomplete linkage disequilibrium (LD) between the analysed SNPs and the causal variants (Yang *et al.*, 2010). Larger sample sizes, refined phenotypes, more densily genotyped SNPs as well as improved statistical methods might help to find the missing heritability.

As follow-up of a detected association between a SNP and a phenotype, the gene underlying the observed association has to be determined. Here, biological knowledge about genes in the considered genomic region and the analysis of their transcript can bring further insight. Moreover, the causal variant underlying the association does not have to be among the significant SNPs as only a fraction of the existing SNPs was analysed. Thus, the genomic region has to be sequenced within fine-mapping approaches. In addition to the determination of the causal genetic variant, functional studies are needed to reveal biochemical mechanisms influencing the observed association (McCarthy *et al.*, 2008). These effords can be complemented by *in silico* analyses of metabolomics and proteomics data (Plomin *et al.*, 2009). So far, GWAS are only a first step in the investigation of genetical and biochemical mechanisms of a complex disease and their risk factors. The hypotheses generated by GWAS together with the new candidate genes have to be further investigated.

1.2 Metabolomics

Metabolomics is the rapidly evolving field of measuring endogenous organic compounds of a cell or body fluid. It is estimated that the human metabolome, which is defined as the complete set of all low-molecular weight molecules, comprises at least 3000 different metabolites of various biochemical classes such as amino acids, lipids, sugars or carnitines (Koal and Deigner, 2010). Metabolites are influenced by genetic factors but also by environmental factors and are involved in many biochemical processes of the cell. Therefore, the analysis of metabolites can reveal insight on functional alterations in the cell and help to detect latent connections between different diseases (Holmes *et al.*, 2008; Barderas *et al.*, 2011). Furthermore, metabolomics is a highly sensitive technique for functional analyses because metabolites are downstream products of genetic and proteomic processes.

As a result, changes in the organism are amplified in the metabolome compared to the genome or proteome. These characteristics make metabolomics a promising tool in the search for biomarkers which help to detect a disease early, to improve the disease prognosis, to evaluate drug toxicity or to develop therapeutics (Nicholson and Lindon, 2008; Nagrath *et al.*, 2011). For example, metabolomics plays an emerging role in the field of cancer diagnostics and therapeutics, especially when early detection is difficult such as for kidney cancer (Nagrath *et al.*, 2011). The search for biomarkers is also upcoming in cardiovascular diseases. However, only minimal improvements over conventional factors were achieved, so far (Barderas *et al.*, 2011). Furthermore, ratios between metabolite concentrations (metabolite ratios) are used in addition to raw metabolite concentrations in the search for biomarkers, e.g. in systematic screens for genetic deficiencies in newborns. An example are elevated concentrations of acylcarnitine ratios which allow to detect medium-chain acyl-coenzyme A dehydrogenase deficiency (Maier *et al.*, 2005). Another example is the phenylalanine to tyrosine ratio which is used to identify heterozygous carriers of phenylketonuria risk alleles (Hsia, 1958). Metabolite ratios are also used as biomarkers for detecting specific exposures. For instance, the urinary hydroxyproline to creatinine ratio was proposed as an indicator for personal exposure to nitrogen dioxide (Yanagisawa *et al.*, 1986).

The measurement of metabolites reveals a snapshot of the current state of cells in the analysed biospecimen. Predominantly, metabolomics analyses are based on blood and urine as these biospecimens are easy to obtain. In principle, there are two analysis strategies to measure metabolites. Whilst the non-targeted approach aims at measuring all metabolites of a biospecimen, the targeted approach focuses on the quantification of selected metabolites. The most accepted high-throughput methods to measure metabolites are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy (Malet-Martino and Holzgrabe, 2011). Among the different NMR methods, mainly ^{1}H-NMR is used, which detects hydrogen atoms in metabolites. NMR methods have the advantage that the analyte does not require any treatment prior to analysis. In contrast, MS has to be coupled to separation techniques, e.g. gas chromatography (GC) or liquid chromatography (LC) but is usually more sensitive than NMR (Nicholson and Lindon, 2008). When using GC/MS, the analyte has to be volatile and thermally stable and sometimes requires a derivatisation step. Among others, fatty acids, organic acids and sugars can be measured with GC/MS very well. If a derivatisation is not possible or if the metabolites are not volatile, LC/MS

can be applied (Barderas *et al.*, 2011). In some cases, tandem MS (MS/MS) is applied which consists of multiple MS steps with a fragmentation step in between. The use of MS/MS facilitates the identification of the measured molecules (Horgan *et al.*, 2008). All together, the combination of different measurement techniques is essential to gain the most comprehensive insight into the metabolome.

1.3 Genome-wide association studies with metabolomics

As metabolites are downstream products of genetic as well as proteomic processes, metabolites are closer connected to genetics in contrast to most of the other analysed phenotypes. The investigation of the genetical basis of metabolites can be achieved through the conduction of metabolomics GWAS.

The first GWAS with metabolomics was done by Gieger *et al.* (2008). They analysed 363 metabolites measured in 284 serum samples. The evaluated metabolite data set comprised not only of lipids but also of amino acids, acylcarnitines and sugars. As initial analysis, a GWAS was conducted for each of the measured metabolite concentrations. Since this analysis did not reveal a significant association, GWAS of metabolite ratios were calculated in a follow-up step. It is considered that the analysis of metabolite ratios increases the statistical power, because systematic experimental errors that are common to the tested metabolite pair are cancelled out, e.g. variance in sample dilution due to pipetting inaccurancies. Furthermore, metabolite ratios can serve as proxies for enzymatic reaction rates for closely biologically connected metabolites. Thus, it is expected that associations with genes encoding enzymes are stronger for metabolite ratios than for single metabolite concentrations. As a result of the metabolite ratio analysis, associations with the *FADS* cluster (fatty acid desaturase) and the *LIPC* locus (hepatic lipase) were discovered, among others. In addition to further insight into biochemical mechanisms, it was also observed that the use of metabolite ratios strengthens the association of multiple orders of magnitudes compared to single metabolite concentrations. After increasing the sample size to 1809 participants, the metabolite concentration and metabolite ratio GWAS were repeated and 15 loci were discovered of which nine could be replicated (Illig *et al.*, 2010). Many of the detected loci were located near enzyme-coding or

solute-carrier coding genes whose proteins match the associated metabolic trait. Hence, these 15 loci helped to discover various processes of the human metabolism.

In the meantime, several metabolomics GWAS were conducted. Examples for lipid based metabolites are GWAS which focus on phospho- and sphingolipids (Hicks *et al.*, 2009; Demirkan *et al.*, 2012), different polyunsaturated fatty acids (Tanaka *et al.*, 2009b; Lemaitre *et al.*, 2011) and lipoprotein subfractions (Chasman *et al.*, 2009). In addition, a GWAS for metabolites measured in human urine samples was also carried out (Suhre *et al.*, 2011b). This GWAS focused on the detoxification capacity of the human body and revealed loci associated with chronic kidney disease and coronary artery disease, among others.

In the first metabolomics GWAS, the capability of metabolite ratio analyses was discovered (Gieger *et al.*, 2008). Whilst in some GWAS all possible pair-wise metabolite ratios were analysed in a hypothesis-free approach, others focused on biologically relevant metabolite ratios. In order to quantify the strengthening in association when analysing metabolite ratios as compared to single metabolite concentrations, the p-gain was introduced. The p-gain for the metabolite ratio M_1/M_2 at a genetic locus X is defined as

$$\text{p-gain}\left(\frac{M_1}{M_2}\middle| X\right) := \frac{\min(\text{p-value}(M_1|X), \text{p-value}(M_2|X))}{\text{p-value}(\frac{M_1}{M_2}|X)},$$

with 'p-value$(M_i|X)$' representing the p-value of the association between the genetic locus X and metabolite M_i, $i = 1, 2$. So far, only a rule of thumb was applied for determination of relevance of the p-gain because the specification of the distribution of the p-gain and therefore of critical values is pending.

As the genetical analysis of metabolites is an evolving field, only some easily obtained gains were achieved, so far. The already measured metabolite concentrations together with their ratios have to be investigated more accurately using statistical and biochemical methods. Moreover, with the development of technologies to measure additional metabolites, analyses of these metabolites will bring further insight into the human metabolism and disease causing mechanisms.

2. Aims of this thesis

Hitherto, GWAS of metabolites were the chosen method to incorporate large-scale metabolomics data in the GWAS approach as well as to investigate the genetical basis of metabolites. Instead of using a hypothesis-free approach it is also possible to conduct a candidate locus approach using current knowledge for the selection of genetic loci. Thus, there are two procedures for the incorporation of metabolomics data in the GWAS approach:

a. Using metabolites for an in-depth analysis of genetic candidate loci which have already been discovered in GWAS of clinically relevant phenotypes.

b. Discovering new genetic loci through conduction of GWAS with metabolites followed by an analysis of these loci together with clinically relevant phenotypes.

In the first procedure (a) metabolites can reveal functional insight into the mechanisms underlying an observed association between a genetic locus and a phenotype. In contrast, in the second procedure (b) metabolites are used to detect novel genetic loci. These detected loci can then serve as candidate loci for clinically relevant phenotypes in order to gain greater insight into disease causing mechanisms. In the following, we refer to the first procedure (a) as candidate locus approach and to the second procedure (b) as metabolomics GWAS approach.

The first aim of this thesis is to compare the two procedures regarding their objectives, advantages, limitations and feasibility. Therefore, we apply the candidate locus approach to 15 lipoprotein subfractions which we analyse together with 95 lipid loci that were discovered in serum lipid GWAS. In addition, we conduct GWAS of over 250 metabolite concentrations and all pair-wise metabolite ratios covering about 60 biochemical pathways as application of the metabolomics GWAS approach. After a presentation of the findings of both applications in Chapter 4 (Results), we compare the procedures in Chapter 5 (Discussion and Conclusion).

For the two procedures, it is possible to analyse not only metabolite concentrations but also pair-wise metabolite ratios. In this case, the p-gain should be applied as an objective measure. Since the distribution of the p-gain is not specified, so far, our second aim is to improve the metabolite ratio analysis through a statistical exploration of the p-gain. In detail, we determine the distribution of the p-gain and derive critical values for different settings of correlations among the metabolic traits. In addition, we show the power of the p-gain approach at the example of the application of the metabolomics GWAS. Therefore, we conduct a pathway enrichment analysis where we compare for metabolite ratios with significant p-gain the membership to a common pathway with that of metabolite ratios with non-significant p-gain. In Chapter 5 (Discussion and Conclusion), we consider the implications of the statistical exploration of the p-gain for the two procedures and the presented applications.

3. Material and methods

In order to address these objectives, we based our analyses on two different sets of metabolites and a total of four different studies. The metabolites and studies are described in the first Section of this Chapter followed by separate methods Sections for each of the two procedures as well as for the statistical p-gain examination.

3.1 Material

3.1.1 Metabolites

For the candidate locus approach, we used 15 lipoprotein subfractions to further characterise 95 lipid loci whereas we used a broad spectrum of metabolites covering different biochemical pathways in the application of the metabolomics GWAS approach. These sets of metabolites were measured using two different technologies.

The lipoprotein subfraction distribution was assessed by NMR spectroscopy and carried out at LipoFIT GmbH, Regensburg, Germany. The technology has been patented (Huber *et al.*, 2005, 2011a,b). Briefly, diffusion-weighted NMR spectra of blood plasma were recorded on a Bruker 600 MHz spectrometer Avance IIplus which revealed characteristic overall profiles of the lipoprotein signals. Using the LipoFIT proprietary software, the regions of the spectra ranging from 0.6 to 1.5 ppm were decomposed into a set of 15 lipoprotein subfractions termed L1-L15 that are characterised by different diffusion constants. The subfractions were defined by LipoFIT in such a way that the corresponding diffusion constants agreed with the presumed particle sizes given in Table A.1 in the Appendix and correspond essentially to small, medium, large and very large high density lipoprotein (HDL) (L1-L4), very small, small, medium, large and very large low density lipoprotein (LDL) (L5-L9), intermediate density lipoprotein (IDL) (L10), small and large very low density lipoprotein (VLDL) (L11 and L12), remnants (L13) and small

and large chylomicrons (L14 and L15) (Linsel-Nitschke *et al.*, 2009). Since for the calculation of particle numbers from the NMR data one has to make additional assumptions about the shape, density and composition of these particles which may bias the statistical analysis, we used the concentrations c_i of methyl groups from cholesterol and fatty acids in the different particle classes Li (i=1, ..., 15), which can be directly measured by NMR.

For the application of the metabolomics GWAS approach, we evaluated metabolites measured by Metabolon, an US commercial supplier of metabolic analyses. For the metabolic profiling, they used two separate ultrahigh performance LC/MS/MS injections and one GC/MS injection per sample (Evans *et al.*, 2009). "The resulting (...) data were searched against a standard library generated by Metabolon (...) [which, AK.P.] allowed for the identification of the experimentally detected molecules (...)" (Suhre *et al.*, 2011a). In total, more than 250 metabolites were profiled, covering over 60 biochemical pathways of the human metabolism. The super pathways to which these metabolites belong to are lipids, carbohydrates, amino acids, nucleotides, peptides, xenobiotics, cofactors and vitamins, among others. A full list of the measured metabolites is given in Table A.2 in the Appendix.

3.1.2 Studies

The Cooperative Health Research in the Region of Augsburg (KORA) study is a series of independent, population-based epidemiological surveys and follow-up studies of participants living in the region of Augsburg, Southern Germany (Wichmann *et al.*, 2005). All participants gave signed informed consent and are residents of Germany with a German nationality identified through registration. The Bayerische Landesärztekammer has approved the studies. For most analyses of this thesis, about 1800 samples of the follow-up study KORA F4 ($2006-2008$) of the KORA S4 survey ($1999-2000$) were evaluated. Within the KORA F4 study, 1814 randomly selected participants were genome-wide genotyped using the Affymetrix GeneChip array 6.0. Genotypes were determined using the Birdseed2 clustering algorithm and imputed using IMPUTE v0.4.2 (Howie *et al.*, 2009) based on HapMap II. The blood samples which were used for the measurement of the metabolites were collected between 2006 and 2008 during the KORA F4 examinations. "To avoid variation due to circadian rhythm, blood was drawn in the morning between 8:00 a.m. and 10:00 a.m. after a period of (...) overnight fasting. (...) [One part of

the blood, AK.P.] was drawn into serum gel tubes, gently inverted twice and then allowed to rest for 30 min at room temperature (18 °C − 25 °C) to obtain complete coagulation. The material was then centrifuged for 10 min (2,750 g at 15 °C). Serum was divided into aliquots and kept for a maximum of 6 h at 4 °C, after which it was deep-frozen to −80 °C until analysis" (Suhre *et al.*, 2011a). These serum samples were used for the metabolite measurements at Metabolon. Another part of the blood was drawn into ethylene diaminetetraacetic acid (EDTA) tubes, gently inverted two times and left on the Sarstedt Universal mixer less than 5 min to avoid mechanical hemolysis, followed by centrifugation for 10 min and 2,750 g at 15 °C. Thereafter, plasma was separated, divided into 200 μl aliquots and kept at 4 °C, after which it was deep-frozen to −80 °C. After less than two weeks, plasma was stored in the gaseous phase of liquid nitrogen (−196 °C). Following the transport on dry ice to Regensburg for lipoprotein subfraction measurement it was kept deep-frozen at −80 °C for two months. Then, plasma was thawed and immediately analysed. Serum lipids were measured on fresh samples using the Dimension RxL (Dade Behring). Total cholesterol (TC) was determined by cholesterol esterase method (CHOL Flex, Dade-Behring, cholesterol oxidase-p-aminophenazone (CHOD-PAP) method), HDL cholesterol (HDL-C) using the AHDL Flex (Dade-Behring, CHOD-PAP method after selective release of HDL-C), LDL cholesterol (LDL-C) using the ALDL Flex (Dade Behring, CHOD-PAP method after colourless usage of all non-LDL-C) and triglycerides (TG) were measured using a TGL Flex (Dade Behring, enzymatic colorimetric test, glycerol phosphate oxidase-p-aminophenazone (GPO-PAP) method). In the following, we refer to serum lipids as the four traits HDL-C, LDL-C, TG and TC whereas we refer to lipoprotein subfractions as L1-L15, which were measured in plasma.

The application of the candidate locus approach to lipoprotein subfractions was done on 1791 samples of the KORA study. For replication of the results, data from 15 samples of the Human Metabolome (HuMet) study as well as from 1940 samples of the Genetic Regulation of Arterial Pressure of Humans in the Community (GRAPHIC) study was evaluated.
The HuMet study is a highly controlled human trial of 15 young and metabolically healthy men which were recruited with a very narrow age range and normal BMI at the Human Study Center in Weihenstephan, Germany (Krug *et al.*, 2012). For a characterisation of the lipoprotein subfractions, data of the lipid tolerance test of the HuMet study was evaluated. The oral lipid tolerance test drink consisted of a 3 : 1 mixture, containing three parts Fresubin® Energy Drink chocolate (Fresenius

Kabi, Bad Homburg, Germany) and one part Calogen® (Nutricia, Zoetemeer, Netherlands). Calogen® is a fat emulsion containing 50 g of long-chain TG per 100 ml. The test drink was served at room temperature at 8:00 a.m. after an overnight fast for ingestion within 5 min. Plasma collections were performed after 0 min, 30 min, 60 min, 90 min, 120 min, 180 min, 240 min and 300 min after the lipid ingestion. For comparison, fasting samples were taken on three days at 8:00 a.m. The second fasting sample was taken four weeks after the first fasting sample. The third fasting sample was taken 24 h after the second fasting sample. This trial was approved by the ethical commission of the Technische Universität München (#2087/08). Blood samples were collected into 9 ml EDTA K_2-Gel tubes (Sarstedt, Nümbrecht, Germany). EDTA-tubes were immediately centrifuged at 3,000 g for 10 min at 20 °C. Plasma was aliquoted by an automatic pipette and was immediately deep-frosted on dry ice and stored at −80 °C until analysis, except for the duration of the transport to Regensburg on dry ice.
The GRAPHIC study was used to replicate the findings of the genetic association analysis. For the GRAPHIC study 2024 individuals from 520 nuclear families of white European origin from Leicestershire in the United Kingdom were recruited. The details of recruitment, phenotyping and sample analysis have been reported by Tomaszewski *et al.* (2010). In brief, for families to be included both parents had to be aged 40 to 60 with two offspring aged 18 or over, with all members agreeing to take part in the study. A standardised questionnaire was used to obtain a comprehensive medical history from participants followed by physical examination, anthropometric measurements, clinic and 24 h ambulatory blood pressure monitoring. The standard biochemistry measurements including HDL-C and TC were performed on non-fasting serum samples using enzymatic assays in an Olympus AU5430 analyser (Samani *et al.*, 2008). Genotypes were determined for the GRAPHIC study using the Illumina HumanCVD BeadChip array (Tomaszewski *et al.*, 2010).

The application of the metabolomics GWAS was done on 1768 KORA samples as well as on 1052 samples of the TwinsUK cohort. "The TwinsUK cohort is a British adult twin registry (...). These unselected twins were recruited from the general population through national media campaigns and were shown to be comparable to age-matched population singletons in terms of disease-related and lifestyle characteristics" (Suhre *et al.*, 2011a; Andrew *et al.*, 2001). Written informed consent has been given by all participants and the study has been approved by the Guy's and St. Thomas' Hospital Ethics Committee. "Blood samples were taken after at

least 6 h of fasting. The samples were immediately inverted three times, followed by 40 min of resting at 4 °C to obtain complete coagulation. The samples were then centrifuged for 10 min at 2,000 g. Serum was removed from the centrifuged brown-topped tubes as the top, yellow, translucent layer of liquid. Four aliquots of 1.5 ml were placed into skirted microcentrifuge tubes and then stored at −45 °C until sampling" (Suhre *et al.*, 2011a). Genotyping of the TwinsUK data set was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M) (Richards *et al.*, 2008; Soranzo *et al.*, 2009). The Illluminus calling algorithm (Teo *et al.*, 2007) was used to assign genotypes. After extensive quality control, the data sets were merged and imputed using IMPUTE v2 (Howie *et al.*, 2009) with HapMap II as well as an own panel as reference.

The statistical analyses of the HuMet, GRAPHIC and TwinsUK cohorts were done by investigators of the studies.

3.1.3 Genotypes

For the application of the candidate locus approach to lipoprotein subfractions, 101 SNPs at 95 lipid loci published by Teslovich *et al.* (2010) were extracted from the imputed genotypes of the KORA study (Table A.3). For replication, the same SNPs or SNPs in LD of more than 0.5 were seleced from the GRAPHIC study.
The metabolomics GWAS of the second approach were based on all genotyped SNPs of the KORA and TwinsUK studies. For fine-mapping of interesting genomic regions, a detailed analysis was conduced using imputed genotype data of the two cohorts.

3.2 Methods

3.2.1 Application of the candidate locus approach

For the evaluation of the lipoprotein subfractions together with the 95 lipid loci, we first characterised the lipoprotein subfractions using serum lipids. This was necessary since we used the concentrations c_i of the lipoprotein subfractions Li (i= $1, \ldots, 15$) and not further derived values such as size or density. Therefore, we conducted a cluster analysis of the lipoprotein subfractions together with the

serum lipids. Moreover, plasma samples from the HuMet study for which measurements were available at three fasting time points as well as at seven time points during a lipid tolerance test were analysed to further characterise the lipoprotein subfractions. After this exploratory work, we calculated associations between the 15 NMR-measured lipoprotein subfractions and 101 genetic variants within 95 lipid loci identified in GWAS (Teslovich *et al.*, 2010, Table A.3). Additionally, we tested the increase in information when analysing lipoprotein subfractions compared to serum lipids using the p-gain approach. The inter-relationship among the lipoprotein subfractions and the associations of the lipid loci were analysed in 1791 plasma samples of the KORA study. The replication of the significant results of the lipid loci analysis was conducted in 1940 samples of the GRAPHIC study.

Data transformation. For the statistical analysis, all serum lipid and lipoprotein subfraction values were naturally log-transformed to achieve normality. Summary statistics for serum lipids and lipoprotein subfractions are combined in Table A.4 in the Appendix.

Characterisation of lipoprotein subfractions

Correlation matrix. We used the 'cor' function implemented in the R-Project Environment (R Development Core Team, 2010) to calculate the Pearson correlation matrix of lipoprotein subfractions and serum lipids for all pair-wise complete observations. Furthermore, we conducted a linear regression analysis for each serum lipid separately with all lipoprotein subfractions as well as age and sex as explaining variables to calculate the proportion of variance of the serum lipids which is explained by the subfractions, age and sex.

Cluster dendrogram. In order to visualise the correlation structure within the lipoprotein subfraction data set, we used an unrooted phylogeny tree where the length of each branch represents the distance between variables. This tree was plotted by using the package 'ape' (Paradis *et al.*, 2004) within the R-Project Environment. The distance measure was based on the correlation between two variables and for the clustering of the lipoprotein subfractions the average linkage method was used. In addition, we applied a bootstrap method implemented in the 'pvclust' package (Suzuki and Shimodaira, 2006) of the R-Project Environment with 10 000 bootstrap replications. In order to measure the confidence of each branch, we used the approximately unbiased (AU) probability, which is more accurate than the bootstrap probability (Shimodaira, 2002). The AU probability

was calculated on multiscale bootstrap resamplings. Beside AU probabilities, we also calculated standard errors to evaluate the confidence of each branch. High AU probabilities and low standard errors indicate a strong support for a branch. For the 15 HuMet samples were multiple measurements at fasting time points as well as during a lipid tolerance test available. Aiming at illustrating the variation between variables and not within individuals for the fasting dendrogram, the clustering of the lipoprotein subfractions was based on average values of multiple measurements from a participant. For the cluster plot of the lipoprotein subfractions during the lipid tolerance test, we aimed to illustrate the variation over the time, so average values of the measurements retained at one time point from all participants were calculated. In a second step, we incorporated the serum lipids in the cluster analysis of KORA samples to classify the lipoprotein subfractions in a natural way.

Development plots. Time dependent graphs were plotted for each cluster to visualise the development of the subfractions during the lipid tolerance test. In order to visualise the change of the subfractions in comparison to the measurement at the starting time point, log-fold changes were used. A fold change is the ratio of a measurement at a certain time point to the measurement at the starting time point. Through calculation of the logarithm ($\log_{10}$), the y-axis represents the change with positive values as increase and negative values as decrease.

Association with 95 lipid loci

Discovery. We analysed in KORA the 101 candidate SNPs described by Teslovich *et al.* (2010) to genetically characterise the lipoprotein subfractions. Therefore, we used the software QUICKTEST (Johnson and Kutalik, 2008) with an additive linear model with age and sex as covariates. In order to correct for multiple testing, we applied Bonferroni correction for the 101 candidate SNPs and 15 lipoprotein subfractions, i.e. p-value $< 3.3 \times 10^{-5} = \frac{0.05}{(101 \cdot 15)}$. Additionally, we calculated p-gain values to test the increase in information due to analysing lipoprotein subfractions compared to serum lipids. Hence, we defined the p-gain as

$$\begin{aligned}&\text{p-gain(lipoprotein subfraction)}\\ &= \frac{\min(\text{p-value(HDL-C)}, \text{p-value(LDL-C)}, \text{p-value(TG)}, \text{p-value(TC)})}{\text{p-value(lipoprotein subfraction)}}.\end{aligned}$$

We defined a SNP as clearly stronger associated with a subfraction than with a serum lipid if the p-gain for a lipoprotein subfraction at a SNP was greater than 15. Finally, the explained variance of a SNP was calculated as the difference be-

tween the explained variance of a linear model with SNP, age and sex as explaining variables and of a linear model with only age and sex as explaining variables.

Replication. In silico replication of the significant associations in the KORA study was conducted in the GRAPHIC study. The analysis of association was carried out using generalised estimation equations with exchangeable correlation structure to account for familial correlations, adjusted for age, age^2 and sex under an additive model of inheritance (Tomaszewski *et al.*, 2010). We applied a Bonferroni correction for the significant SNP - lipoprotein subfraction associations to correct for multiple testing.

3.2.2 Application of the metabolomics GWAS approach

For the GWAS of the metabolites, we decided to analyse not only all metabolite concentrations ($N = 276$ in KORA) but also all pair-wise metabolite ratios ($N = 37\,179$ in KORA), in total 37 455 metabolic traits in KORA, since the analysis of metabolite ratios showed good results in Gieger *et al.* (2008) and Illig *et al.* (2010). Due to the increased computational and data storage burden, we conducted a stepwise approach. First, we performed all metabolite concentration and metabolite ratio GWAS on genotyped SNPs. Then, we selected promising signals between genomic regions and metabolic traits and repeated the association analysis on genotyped and imputed SNPs of these regions. For loci which were significant in this fine-mapping analysis, we specified candidate genes and clinically relevant phenotypes which were reported to be associated with these loci. As a follow-up analysis, we calculated associations between the metabolic traits and selected clinically relevant phenotypes.

Quality control of metabolites and genotypes. For quality control of the metabolomics data set, all data points with a distance of more than three standard deviations to the mean of the metabolic traits were excluded. Moreover, only metabolic traits with at least 300 non-missing values were analysed. In total, 276 metabolite concentrations and 37 179 metabolite ratios were available in KORA whereas in TwinsUK 258 metabolite concentrations and 32 499 metabolite ratios were available. A test of normal distribution for the metabolic traits showed that for more cases the $\log_{10}$-transformed values were closer to the normal distribution than the untransformed values. Therefore, $\log_{10}$-transformation was applied to all

metabolic traits. Moreover, testing ratios between two metabolite concentrations a and b should be independent of their order. This is achieved when analysing log-scaled metabolite ratios due to the property $\log(a/b) = -\log(b/a)$. This also halves the multiple testing burden.
As quality control of the genotypes, we excluded all SNPs with a call rate less than 95 % and a p-value $< 10^{-6}$ for deviation from the Hardy-Weinberg equilibrium. In total, about 655 000 autosomal SNPs were included in the GWAS of the KORA study and about 535 000 autosomal SNPs in the GWAS of the TwinsUK study.

Metabolomics GWAS

GWAS and meta-analysis. The metabolomics GWAS were carried out on genotyped SNPs using an additive linear regression model for all metabolic traits. We adjusted for age, sex and family structure. For the GWAS, the software PLINK v1.06 (Purcell *et al.*, 2007) and SNPTEST (Marchini *et al.*, 2007) were used in KORA whereas Merlin (Abecasis *et al.*, 2002) which accounts for family structure was used in the TwinsUK study. In order to measure the strengthening in association when analysing a metabolite ratio compared to the single metabolite concentrations, the p-gain approach was applied. Furthermore, we calculated the inflation factor λ and plotted quantile-quantile plots to check for inflation of summary statistics which can reflect population stratification in the analysed sample or an unappropriate statistical model (Devlin and Roeder, 1999; De Bakker *et al.*, 2008). After this initial GWAS on genotyped SNPs, we selected the genomic regions and metabolic traits which had an association p-value $< 10^{-6}$ in both cohorts or a p-value $< 10^{-3}$ in one and a p-value $< 10^{-9}$ in the other cohort for further analysis. Additionally, for metabolite ratios we required the p-gain to be larger than 250. For each of these genomic regions, associations were calculated for both cohorts between the genotyped and imputed SNPs of the genomic region and the selected metabolic traits. Afterwards, the results were meta-analysed using the fixed-effects inverse variance method (De Bakker *et al.*, 2008). The combination of SNP and metabolic trait that yielded to the smallest p-value in this meta-analysis was finally selected. In the following, we refer to the SNP with the smallest p-value in the meta-analysis as lead SNP for the genomic region.

Correction for multiple testing. A conservative Bonferroni correction for multiple testing was applied using the KORA study as a reference. The nominal significance level of 5 % was corrected for tests on 655 658 SNPs and 37 455 metabolic traits. This resulted in a Bonferroni corrected level of significance of 2.0×10^{-12}.

For metabolite ratios, it was also required that the p-gain has to be larger than 250 which is approximately the number of tested metabolite concentrations.

Follow-up analysis of GWAS results

Candidate gene selection. Using knowledge about the function of genes which are located near the lead SNP and about the biochemical characteristics of the associated metabolic traits, we identified a single most likely candidate gene in many cases.

Overlap with published associations. For each locus, SNPs were identified which were previously reported to be associated with clinically relevant phenotypes. These SNPs were required to have an LD of more than 0.8 with the lead SNP. This search was done using the catalogue of published GWAS (Hindorff *et al.*, 2011).

Associations with clinically relevant phenotypes. For selected loci we further tested the association between a metabolic trait and a clinically relevant phenotype through calculation of linear regression models. One tested clinically relevant phenotype was the estimated glomerular filtration rate (eGFR) which is defined as

$$\text{eGFR} = 175 \times \text{scr}^{-1.154} \times \text{age}^{-0.203} \times 1.212 \text{ (if black)} \times 0.742 \text{ (if female)}$$

with scr the serum creatinine measurement in mg/dl (Levey *et al.*, 2007). Another clinically relevant phenotype which we analysed in the follow-up analysis was hypertension. We defined a sample as hypertensive if the systolic blood pressure was higher than 190 mmHg and the diastolic blood pressure was higher than 90 mmHg or if the sample was on anti-hypertensive medication.

3.2.3 Statistical exploration of the p-gain

In order to statistically explore the p-gain, we derived critical values through determination of the distribution of the p-gain. In case of uncorrelated metabolic traits, the distribution can be calculated. For the other cases, we conducted a simulation approach. In addition, we investigated the characteristics of the p-gain in the situation of Bonferroni correction for multiple testing as well as the depencence of observed p-gain values on the sample size. Finally, we illustrated the power of the p-gain approach by investigating the enrichment for common pathways among

metabolite ratios with significant p-gain. The analysis of the dependence of observed p-gain values on the sample size as well as the pathway enrichment analysis were based on the concrete example of the application of the metabolomics GWAS (Chapter 3.2.2 and 4.2).

Density of p-gain

For uncorrelated metabolic traits (calculation). As notation, we used 'p-value($M_i|X$)', short '$P(M_i)$', to reference the p-value corresponding to a test for association between a genetic locus X and the metabolite M_i, $i = 1, 2$. This is often the test of the effect size in a linear regression of a genetic locus X to the metabolite M_i. With this definition, the p-gain for the ratio M_1/M_2 of metabolites M_1 and M_2 at a genetic locus X is defined as

$$\text{p-gain}\left(\frac{M_1}{M_2}\middle|X\right) := \frac{\min(\text{p-value}(M_1|X), \text{p-value}(M_2|X))}{\text{p-value}(\frac{M_1}{M_2}|X)}. \tag{1}$$

We further define the universal p-gain as the ratio of p-values belonging to two uncorrelated metabolic traits:

$$\text{p-gain}_{\text{univ}}\left(\frac{M_1}{M_2}\middle|X\right) := \frac{\text{p-value}(M_1|X)}{\text{p-value}(\frac{M_1}{M_2}|X)},\ \text{cor}(M_1, \frac{M_1}{M_2}) = 0. \tag{2}$$

Critical values of the distribution of the universal p-gain are conservative to the critical values of the distribution of the p-gain given in equation (1) because

$$\text{p-value}(M_1|X) \geq \min(\text{p-value}(M_1|X), \text{p-value}(M_2|X))$$

and therefore

$$\frac{\text{p-value}(M_1|X)}{\text{p-value}(\frac{M_1}{M_2}|X)} \geq \frac{\min(\text{p-value}(M_1|X), \text{p-value}(M_2|X))}{\text{p-value}(\frac{M_1}{M_2}|X)}.$$

The variation of the distribution of the p-gain defined in equation (2) depends on the correlation among M_1 and M_1/M_2. For example, highly correlated metabolic traits contain mainly the same information and have similar p-values in association tests. This results in p-gain values which are close to one. Hence, the variation of the distribution is small. In contrast, weakly correlated metabolic traits contain different information and may have different p-values in association tests. This results in p-gain values distributed broadly around the one. Therefore, assuming $\text{cor}(M_1, M_1/M_2) = 0$, as it was done in equation (2), results in a distribution of the universal p-gain with largest possible variation and leads to the most conservative

critical values. For the universal p-gain, the density can be calculated by using the convolution formula for ratios:

$$f_{\frac{P(M_1)}{P(M_1/M_2)}}(\text{p-gain}) = \int_{-\infty}^{+\infty} |t| \cdot f_{P(M_1)}(\text{p-gain} \cdot t) \cdot f_{P(M_1/M_2)}(t)dt \;\; \forall \text{ p-gain} \in R^+,$$

with $P(M_1)$ and $P(M_1/M_2)$ having a uniform distribution on the interval $[0, 1]$. Transformations lead to

$$\begin{aligned} f_{\frac{P(M_1)}{P(M_1/M_2)}}(\text{p-gain}) &= \int_{-\infty}^{+\infty} |t| \cdot f_{P(M_1)}(\text{p-gain} \cdot t) \cdot f_{P(M_1/M_2)}(t)dt \\ &= \int_0^1 t \cdot f_{P(M_1)}(\text{p-gain} \cdot t)dt \\ &= \begin{cases} \int_0^{\frac{1}{\text{p-gain}}} t\, dt = \frac{1}{2 \cdot \text{p-gain}^2}, & \text{p-gain} \geq 1 \\ \int_0^1 t\, dt = \frac{1}{2}, & 0 < \text{p-gain} < 1. \end{cases} \end{aligned}$$

The corresponding cumulative distribution is

$$F_{\frac{P(M_1)}{P(M_1/M_2)}}(\text{p-gain}) = \int_0^{\text{p-gain}} f_{\frac{P(M_1)}{P(M_1/M_2)}}(t)dt = \begin{cases} 1 - \frac{1}{2 \cdot \text{p-gain}}, & \text{p-gain} \geq 1 \\ \frac{1}{2}\,\text{p-gain}, & 0 < \text{p-gain} < 1. \end{cases}$$

Therefore,

$$\begin{aligned} F_{\frac{P(M_1)}{P(M_1/M_2)}}(\text{p-gain}) = (1 - \frac{\alpha}{B}) &\Leftrightarrow 1 - \frac{1}{2 \cdot \text{p-gain}} = (1 - \frac{\alpha}{B}), \quad \text{if p-gain} \geq 1 \\ &\Leftrightarrow \text{p-gain} = \frac{B}{2\,\alpha}, \quad \text{if } \frac{\alpha}{B} \leq 0.5, \end{aligned}$$

with α/B being the significance level α, Bonferroni corrected for B tests.

For correlated metabolic traits (simulation). To determine the density of the p-gain as defined in equation (1), we assumed a given correlation structure among the metabolic traits. This confers to a correlation structure among p-values corresponding to these metabolic traits. With these correlated p-values the density of the p-gain can be derived. For simulation of the variables with a given correlation structure we chose the 'copula' package (Yan, 2007; Kojadinovic and Yan, 2010) of the R-Project Environment. A copula is a joint probability distribution which one-dimensional marginal distributions are uniformly distributed over the interval $[0, 1]$. It takes the dependency among the marginal distributions into account.

After simulating variables using a copula, we transformed them with an inverse normal transformation to gain normal distributed variables which is essential for linear regressions. To simulate the p-values belonging to these variables, we generated additional variables and conducted linear regressions where these additional variables were the independent and the variables simulated with the copula the dependent variables. The received p-values contain a correlation structure which belongs to the correlation structure of the metabolic traits. Out of these p-values, we calculated the density of the p-gain empirically and derived critical values for given significance levels.

Dependence of p-gain values on sample size

We determined the dependence of p-gain values on the sample size by drawing randomly (with replacement) between 100 and 2000 samples from the KORA data which we used for the application of the metabolomics GWAS (Chapter 3.2.2). For each sample size, we repeated this analysis 1500 times. For all sample subsets we calculated p-gain values. We then determined the median p-gain values as well as the 1^{st} and 3^{rd} quantile of the p-gain values for each sample size.

P-gain and metabolomics pathways

We used the KORA results of the application of the metabolomics GWAS (Chapter 4.2) to analyse the enrichment of pathways for metabolite ratios with a large p-gain. For this analysis, we additionally filtered the GWAS results for minor allele frequency (MAF) greater than 5 % and extracted for each metabolite ratio the SNP with the largest p-gain. As terminology, we defined a metabolite ratio to be on a pathway, whenever both metabolite concentrations of the metabolite ratio belong to the same pathway. For pathway annotations, we applied different mappings such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2006, 2010), the Small Molecule Pathway Database (SMPDB) (Frolkis *et al.*, 2010), levels two and three of the Human Metabolome Database (HMDB) (Wishart *et al.*, 2009), super- and sub-pathways provided by Metabolon (Evans *et al.*, 2009) and Gaussian graphical models (GGMs) (Krumsiek *et al.*, 2011). We coded the pathway information of each data base as one if both metabolite concentrations of a ratio were on the same pathway, else zero. If there was no information available about a metabolite ratio in one mapping, we omitted this particular mapping from the following calculations for this metabolite ratio. With this, we computed a percentage of the mappings which assigned both metabolite concentrations of a ratio to the same pathway and tested the difference

of the mean of pathway allocation for the best 100, 500, 1000 and 1500 metabolite ratios vs. the mean of pathway allocation for all metabolite ratios using a t-test. Additionally, we compared the allocation to a common pathway for all metabolite ratios with a significant p-gain vs. all metabolite ratios with a non-significant p-gain.

4. Results

4.1 Application of the candidate locus approach: Genetic associations with lipoprotein subfractions provide information on their biological nature

Background

At present, 95 associated common variants are reported for HDL-C, LDL-C, TG and TC (Teslovich *et al.*, 2010). These loci explain 10 % − 12 % of the total variance of serum lipids. Extreme levels of serum lipids are a major risk factor for cardiovascular outcomes such as coronary artery disease, myocardial infarction and stroke (Castelli *et al.*, 1977; Castelli, 1996). Whilst the contribution of LDL-C to the development of coronary artery disease is well documented, the role of other lipoprotein fractions (including HDL-C) in atherosclerosis and its clinical manifestations is less well understood (Asztalos *et al.*, 2004; Rader, 2006, 2009). For example, the torcetrapib failure revealed the complexity of the HDL metabolism and implicated that further research on HDL and HDL fractions is needed (Von Eckardstein, 2010). In order to obtain a more detailed view of the lipid metabolism, subfractions of lipoproteins which can be measured using ^{1}H-NMR spectroscopy can be analysed. Using a 400 MHz NMR lipoprotein analyser, Chasman *et al.* (2009) conducted a GWAS of the lipoprotein subfractions with the aim of finding new genetic lipid loci.

The aim of this application is to gain a more in-depth view into biological processes of the lipid metabolism through analysing lipoprotein subfractions together with known genetic lipid loci and to investigate if the analysis of subfractions reveals more and stronger associations with genetic loci than the analysis of serum lipids.

	tree A		tree B		tree C		tree D	
branch	AU	SE	AU	SE	AU	SE	AU	SE
1	0.629	0.009	0.644	0.009	0.751	0.007	0.979	0.001
2	1.000	0.000	0.755	0.009	0.998	0.000	0.996	0.000
3	0.996	0.006	1.000	0.000	0.967	0.002	0.966	0.002
4	0.812	0.008	0.994	0.003	0.995	0.000	0.851	0.006
5	0.899	0.005	0.958	0.004	0.766	0.007	1.000	0.000
6	1.000	0.000	0.721	0.009	0.869	0.005	0.989	0.001
7	0.609	0.009	0.889	0.006	0.893	0.004	1.000	0.000
8	0.817	0.008	0.931	0.005	0.851	0.006	0.994	0.000
9	0.696	0.009	0.687	0.009	0.821	0.006	0.843	0.006
10	1.000	0.000	0.763	0.008	0.936	0.003	0.996	0.000
11	1.000	0.000	0.801	0.007	0.949	0.003	0.990	0.001
12	1.000	0.000	0.790	0.007	0.909	0.004	0.953	0.002
13	1.000	0.000	0.992	0.004	0.870	0.005	0.948	0.002
14			0.995	0.002				
15			1.000	0.000				
16			0.999	0.000				
17			0.999	0.001				

Table 4.1: AU probabilities and standard errors for cluster plots. AU probabilities (AU) and standard errors (SE) of 10 000 bootstrap replications were provided for each branch of the trees of Figure 4.1. High AU probabilities and low standard errors indicate a strong support for a branch. **Tree A** lipoprotein subfractions in KORA, **tree B** lipoprotein subfractions and serum lipids in KORA, **tree C** lipoprotein subfractions in the fasting samples of HuMet and **tree D** lipoprotein subfractions in the HuMet samples during the lipid tolerance test (Petersen *et al.*, 2012).

Results

Inter-relationship of lipoprotein subfractions

In order to be independent of assumptions about the shape of lipoprotein subfractions, we assigned them to the serum lipids in a statistical analysis. First, using linear regressions with all lipoprotein subfractions as explaining variables, we observed that they explained a high proportion of serum lipid variance: 94 % of the variance of TG, 84.6 % of TC, 82.5 % of HDL-C and 75.7 % of LDL-C. To get a more in-depth view into the inter-relationship of lipoprotein subfractions, we conducted a cluster analysis of the subfractions in KORA based on their correlation matrix as a distance measure, followed by bootstrap replications to test the robustness of the clustering. The results of this cluster analysis are displayed in an unrooted tree (Figure 4.1 A). At first observation, the tree indicated that

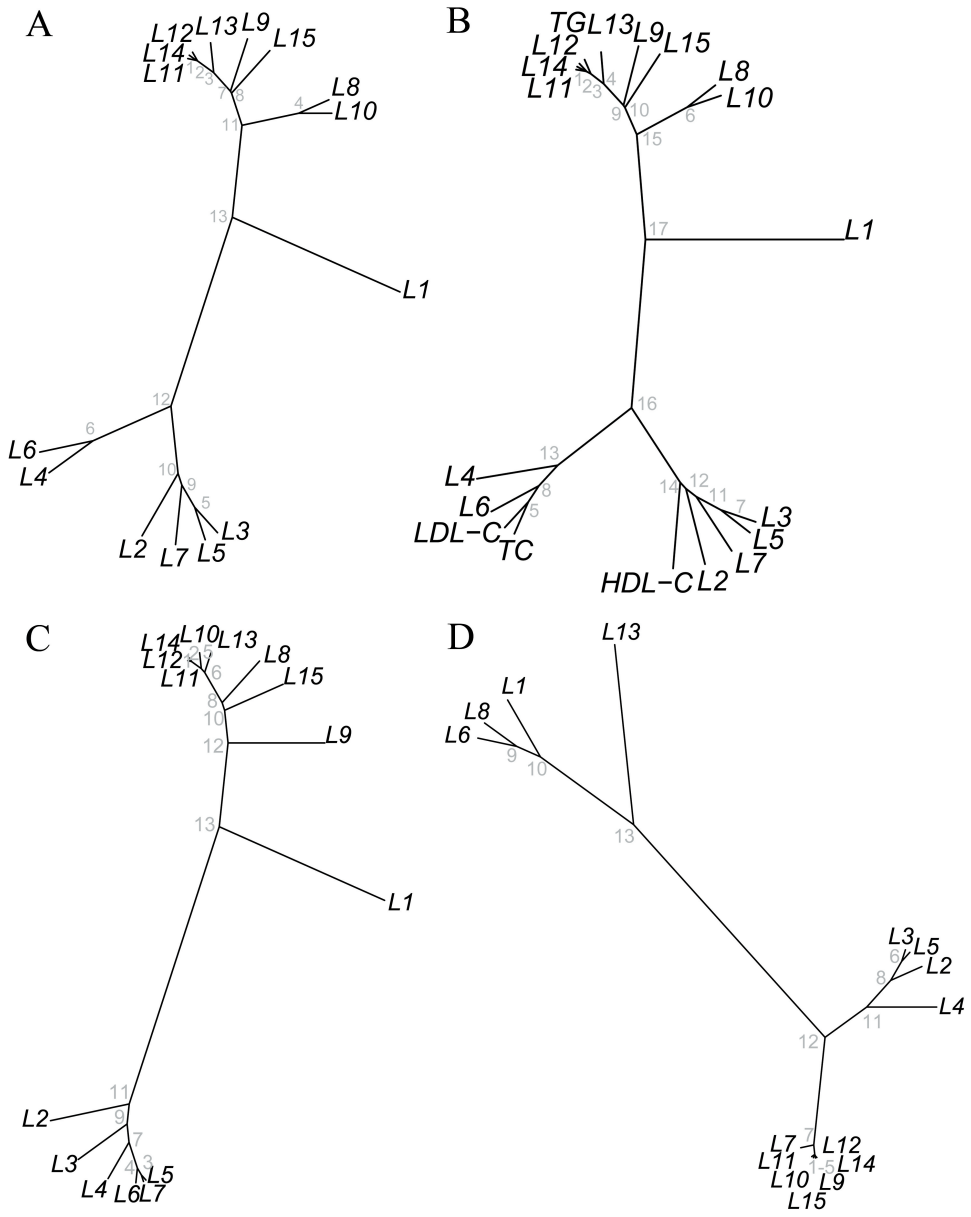

Figure 4.1: Cluster plots of lipoprotein subfractions. The cluster plots of the inter-relationship of the lipoprotein subfractions were displayed in an unrooted phylogeny tree using the correlation between the subfractions as distance measure. The length of a branch represents the distance between the subfractions. Each phylogeny tree was created out of 10 000 bootstrap replications. **A** lipoprotein subfractions in KORA, **B** lipoprotein subfractions and serum lipids in KORA, **C** lipoprotein subfractions in the fasting samples of HuMet and **D** lipoprotein subfractions in the HuMet samples during the lipid tolerance test. In Table 4.1 are for all branches the AU probabilities and the standard errors summarised (Petersen *et al.*, 2012).

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	R^2
HDL-C	−0.11	0.69	0.81	0.39	0.67	0.15	0.51	−0.39	−0.26	−0.52	−0.53	−0.52	−0.54	−0.52	−0.28	0.83
LDL-C	0.29	−0.05	−0.004	0.59	0.29	0.74	0.50	0.35	0.49	0.33	0.29	0.26	0.20	0.21	0.17	0.76
TC	0.20	0.21	0.28	0.70	0.52	0.85	0.68	0.39	0.57	0.37	0.31	0.30	0.20	0.24	0.27	0.85
TG	0.24	−0.23	−0.44	−0.03	−0.28	0.29	−0.10	0.65	0.74	0.88	0.96	0.94	0.89	0.93	0.72	0.94

Table 4.2: Correlations between lipoprotein subfractions and serum lipids. The Pearson correlation coefficient was calculated for each lipoprotein subfraction and serum lipid. The R^2 is the explained variance which was calculated in the linear regression model using all subfractions, sex and age as explaining variables (Petersen *et al.*, 2012).

L1 is separate from the remaining subfractions. Furthermore, two major groups were distinguished: L2-L7 and L8-L15. Each of the two major groups contained two subgroups. In total, we had the following five clusters: (L1), (L2, L3, L5, L7), (L4, L6), (L8, L10) and (L9, L11, L12, L13, L14, L15). For the mentioned intersections, the bootstrap replications revealed an AU probability of one and a standard error of zero, which means that these divisions are absolutely reliable (Table 4.1, tree A). In the next step, we added the serum lipids to the tree to get a lipid-based characterisation of the subfractions. After their inclusion, the main inter-relationships between the subfractions remained unchanged (Figure 4.1 B). We found that HDL-C clustered together with (L2, L3, L5, L7), LDL-C and TC with (L4, L6) and TG with (L9, L11, L12, L13, L14, L15). In the tree with serum lipids, the AU probabilities were smaller than before but the divisions in the mentioned five clusters were still very reliable (Table 4.1, tree B). In order to further characterise the relations between lipoprotein subfractions and serum lipids, we used Pearson correlations. The results revealed that the largest correlation of HDL-C was with L3, of LDL-C and TC with L6 and of TG with L11 (Table 4.2). Surprisingly, lipoprotein subfraction L1 was only weakly correlated with all serum lipids.

Lipoprotein subfractions after nutritional intervention

To investigate whether the clustering of the subfractions was stable after nutritional intervention, we repeated the clustering in plasma samples from the 15 young men of the HuMet study for whom lipoprotein subfraction measurements were conducted at three fasting time points as well as at seven time points during a lipid tolerance test. In the cluster plot of the fasting time points, we replicated the main three clusters: L1, L2-L7 and L8-L15 (Figure 4.1 C). For these intersec-

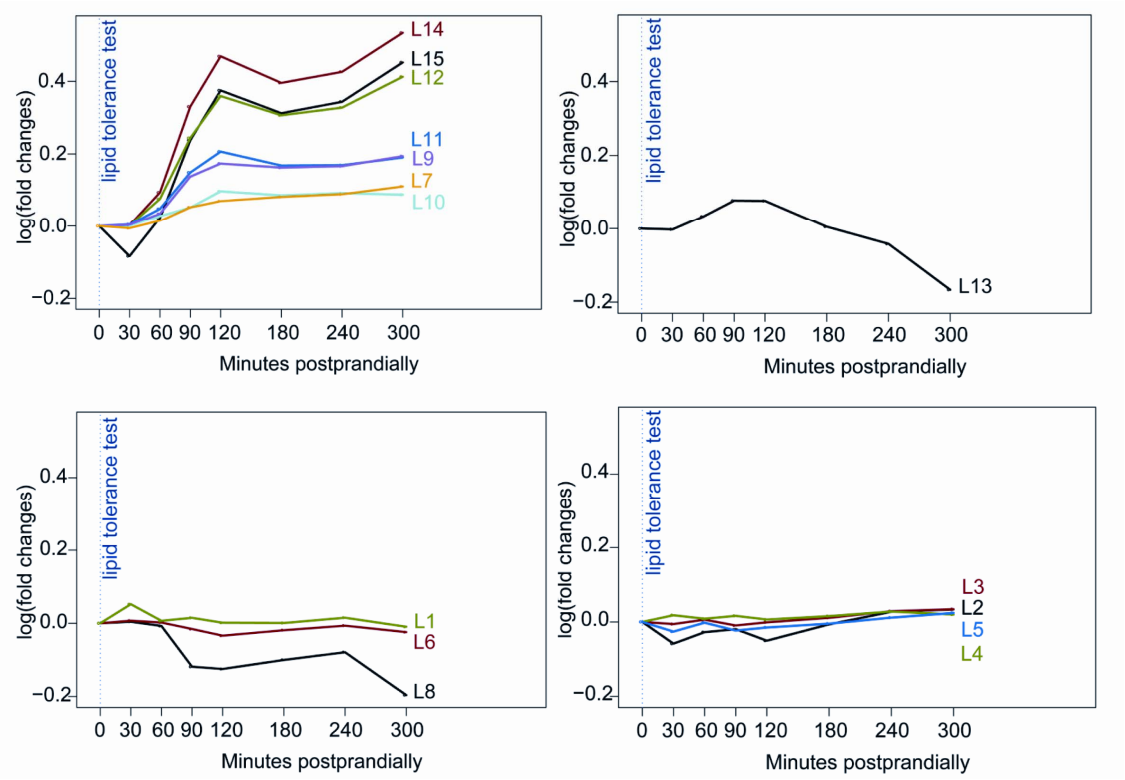

Figure 4.2: Development plots of lipoprotein subfractions during lipid tolerance test. Each panel shows the development of a cluster of the lipoprotein subfractions during the lipid tolerance test (Figure 4.1 D). The x-axis represents the time, the y-axis the log-fold change, which describes the change of a measurement compared to the first measurement (Petersen *et al.*, 2012).

tions, we had reliable AU probabilities and standard errors (Table 4.1, tree C). In contrast to the fasting cluster plot, we observed changes in the clustering of the measurements during the lipid tolerance test (Figure 4.1 D). The lipoprotein subfractions shifted and generated new groups, e.g. (L1, L6, L8). In the fasting cluster plot, L1 was independent of all other subfractions, L6 belonged to the group (L4, L6) and L8 belonged to the group (L8, L10). During the lipid tolerance test, subfraction L13 was independent of the other subfractions. Moreover, subfraction L7 changed from the group (L2, L3, L5, L7) into the group (L7, L9, L10, L11, L12, L14, L15). For the major divisions, we observed again reliable AU probabilities and standard errors (Table 4.1, tree D). Subfractions within each group showed a similar trend during the lipid tolerance test (Figure 4.2). In group (L7, L9, L10, L11, L12, L14, L15), all subfractions increased after 300 minutes but to a different extent. L7 and L10 only increased by about 0.1, whereas L14 increased by

about 0.5. The subfractions (L2, L3, L4, L5) stayed nearly constant during the lipid tolerance test whilst lipoprotein subfraction L13 decreased notably by about 0.2. Thus the lipid tolerance test revealed the different influences of nutritional intervention on lipoprotein subfractions. As a result, subfraction L1 was assorted together with subfractions L6 and L8, which was in contrast to the results of the fasting samples (Figure 4.1).

Proportion of variance explained by genes increases for subfractions
With this mapping of the lipoprotein subfractions, we next tested the association between 101 SNPs and the 15 subfractions using an additive genetic model. In addition to the Bonferroni corrected significance level of 3.3×10^{-5}, we compared the p-value of the subfractions with the p-value of the serum lipids through calculation of a p-gain. Eight of the analysed loci showed significant associations with at least one of the 15 subfractions (Tables 4.3 and 4.4). Moreover, associations with *FADS1-2-3*, *LIPC*, *PLTP*, *APOB* and *APOA1* had relevant p-gains (i.e. p-gain > 15) in KORA whereas for *CETP*, *SORT1* and *GCKR*, use of subfractions did not strengthen the original association. For *FADS1-2-3*, *LIPC*, *CETP*, *PLTP* and *GCKR*, we replicated all significant associations as well as the relevant p-gains in the GRAPHIC study (Table 4.4). For the remaining loci some associations were not significant in GRAPHIC after Bonferroni correction. Nevertheless, the direction of effect at these loci was consistent in KORA and GRAPHIC. In contrast, when analysing associations of serum lipids together with lipid loci, we found that only four loci in KORA were associated (*CETP*, *SORT1*, *GCKR*, *APOA1*). In addition to this, for *FADS1-2-3*, *LIPC*, *PLTP* and *APOB* the explained variance was clearly larger for lipoprotein subfractions than for serum lipids (Figure 4.3). In detail, the explained variances between lipid loci and subfractions were up to 2.3 % (*APOA1* and L8). For serum lipids, we explained up to 1.7 % of the variance (*CETP* and HDL-C). Altogether, the explained variance of the lipoprotein subfractions ranged from 1.5 % (L9) to 4.5 % (L8) and of serum lipids from 1.0 % (TC) to 3.3 % (TG). Summing up these results, we found more significant associations with lipoprotein subfractions and in addition, we could explain more of the variance of lipoprotein subfractions than of serum lipids. As a biological classification of the significant eight genes, Figure 4.4 integrates the genes together with the analysed lipoprotein subfractions in the lipid metabolism. The colours indicate the assignment of the lipoprotein subfractions to the three main clusters: L1, L2-L7 and L8-L15.

Table 4.3: Loci with significant associations with 15 lipoprotein subfractions in KORA. This Table shows the p-values of the eight loci which were associated with at least one of the 15 subfractions in KORA. Results were provided for the 15 subfractions (L1-L15), the serum lipids in KORA (KORA HDL-C, KORA LDL-C, KORA TC, KORA TG) and serum lipids in the global lipids meta-analysis (`http://www.sph.umich.edu/csg/abecasis/public/lipids2010/`) (GLC HDL-C, GLC LDL-C, GLC TC, GLC TG) (Teslovich *et al.*, 2010). The subfractions were ordered according to the clustering of Figure 4.3. P-values highlighted in bold were significant after Bonferroni correction. Detailed results for the significant associations are summarised in Table 4.4 (Petersen *et al.*, 2012).

lipoprotein subfraction	*FADS1-2-3* rs174546	*LIPC* rs1532085	*CETP* rs3764261	*PLTP* rs6065906	*SORT1* rs629301	*GCKR* rs1260326	*APOB* rs1042034	*APOA1* rs964184
L2	0.699	$\mathbf{3.40 \times 10^{-7}}$	1.43×10^{-5}	0.232	0.114	0.477	0.144	0.954
L3	7.61×10^{-3}	$\mathbf{4.22 \times 10^{-7}}$	$\mathbf{3.59 \times 10^{-7}}$	$\mathbf{1.72 \times 10^{-5}}$	0.845	0.420	0.196	0.706
L5	0.019	$\mathbf{5.27 \times 10^{-11}}$	7.32×10^{-4}	0.016	0.622	0.859	0.271	0.822
L7	0.148	$\mathbf{7.28 \times 10^{-10}}$	0.019	0.150	0.050	0.709	0.452	0.966
L4	$\mathbf{1.38 \times 10^{-5}}$	5.25×10^{-5}	5.16×10^{-3}	0.423	3.58×10^{-5}	0.082	0.780	0.865
L6	0.034	3.33×10^{-4}	0.431	0.136	$\mathbf{1.46 \times 10^{-5}}$	1.54×10^{-3}	9.94×10^{-3}	3.04×10^{-4}
L1	0.234	0.015	0.584	$\mathbf{4.86 \times 10^{-7}}$	5.55×10^{-3}	0.170	0.589	0.356
L8	0.913	0.583	0.084	0.806	0.024	$\mathbf{9.25 \times 10^{-6}}$	$\mathbf{1.08 \times 10^{-5}}$	$\mathbf{4.82 \times 10^{-12}}$
L10	0.073	0.788	0.104	0.383	0.017	$\mathbf{3.73 \times 10^{-6}}$	$\mathbf{1.63 \times 10^{-5}}$	$\mathbf{9.47 \times 10^{-11}}$
L9	0.035	0.092	0.311	0.534	5.97×10^{-4}	4.07×10^{-3}	0.385	8.56×10^{-3}
L11	4.56×10^{-3}	0.692	0.107	0.204	7.70×10^{-3}	3.72×10^{-5}	0.025	$\mathbf{6.25 \times 10^{-7}}$
L14	1.72×10^{-3}	0.457	0.089	0.149	0.014	$\mathbf{2.01 \times 10^{-5}}$	0.196	4.49×10^{-5}
L12	5.20×10^{-3}	0.832	0.133	0.169	0.014	$\mathbf{6.88 \times 10^{-6}}$	6.04×10^{-3}	$\mathbf{2.72 \times 10^{-7}}$
L13	2.23×10^{-3}	0.199	0.191	0.316	0.032	1.84×10^{-4}	0.087	$\mathbf{3.16 \times 10^{-7}}$
L15	2.73×10^{-3}	0.419	0.607	0.493	0.067	9.85×10^{-4}	0.731	4.56×10^{-4}
KORA HDL-C	0.907	0.016	$\mathbf{5.69 \times 10^{-9}}$	0.153	0.173	0.525	0.257	0.027
KORA LDL-C	0.594	0.947	0.341	0.759	$\mathbf{1.36 \times 10^{-5}}$	0.214	0.088	0.718
KORA TC	0.416	0.220	0.259	0.444	$\mathbf{8.71 \times 10^{-6}}$	0.020	0.056	0.069
KORA TG	0.022	0.212	0.105	0.077	0.012	$\mathbf{2.54 \times 10^{-7}}$	0.035	$\mathbf{2.09 \times 10^{-8}}$
GLC HDL-C	$\mathbf{2.62 \times 10^{-22}}$	$\mathbf{2.92 \times 10^{-96}}$	$\mathbf{7.10 \times 10^{-380}}$	$\mathbf{1.90 \times 10^{-22}}$	6.19×10^{-8}	0.078	$\mathbf{1.22 \times 10^{-30}}$	$\mathbf{5.21 \times 10^{-47}}$
GLC LDL-C	$\mathbf{1.76 \times 10^{-21}}$	0.852	$\mathbf{1.64 \times 10^{-12}}$	0.297	$\mathbf{9.70 \times 10^{-171}}$	2.33×10^{-4}	$\mathbf{8.32 \times 10^{-25}}$	$\mathbf{1.47 \times 10^{-26}}$
GLC TC	$\mathbf{2.85 \times 10^{-22}}$	$\mathbf{8.83 \times 10^{-20}}$	$\mathbf{6.67 \times 10^{-14}}$	0.970	$\mathbf{5.77 \times 10^{-131}}$	$\mathbf{7.31 \times 10^{-27}}$	$\mathbf{3.71 \times 10^{-18}}$	$\mathbf{6.21 \times 10^{-57}}$
GLC TG	$\mathbf{5.41 \times 10^{-24}}$	$\mathbf{1.78 \times 10^{-11}}$	$\mathbf{6.15 \times 10^{-12}}$	$\mathbf{2.59 \times 10^{-17}}$	0.062	$\mathbf{5.68 \times 10^{-133}}$	$\mathbf{1.36 \times 10^{-45}}$	$\mathbf{6.71 \times 10^{-240}}$

Table 4.4: Detailed results and replication of significantly associated loci. Number of samples (N), coded (effect, minor) allele, non-coded (major) allele, effect size (beta), standard error (SE), p-value of association, increase in the strength of association compared to serum lipids (p-gain) and MAF are reported for KORA and GRAPHIC. Only relevant p-gain values (i.e. p-gain > 15) are provided for KORA. Associations marked with * were replicated in GRAPHIC (Petersen *et al.*, 2012).

					KORA (N=1791)						GRAPHIC (N=1940)						
lipoprotein subfraction	gene	SNP	chr	position	coded/ non-coded	beta	SE	p-value	p-gain	MAF	proxy	coded/ non-coded	beta	SE	p-value	p-gain	MAF	
L2	*LIPC*	rs1532085	15	56470658	A/G	0.083	0.016	3.40×10^{-7}	4.71×10^{4}	0.351	rs4775041	C/G	0.064	0.015	2.39×10^{-5}	1.29×10^{2}	0.307	*
L2	*CETP*	rs3764261	16	55550825	T/G	0.086	0.020	1.43×10^{-5}	—	0.327	—	—	0.061	0.014	9.25×10^{-6}	—	0.321	*
L3	*LIPC*	rs1532085	15	56470658	A/G	0.081	0.016	4.22×10^{-7}	3.80×10^{4}	0.351	rs4775041	C/G	0.067	0.013	5.20×10^{-7}	5.94×10^{3}	0.307	*
L3	*CETP*	rs3764261	16	55550825	T/G	0.099	0.019	3.59×10^{-7}	—	0.327	—	—	0.090	0.013	2.08×10^{-11}	—	0.321	*
L3	*PLTP*	rs6065906	20	43987422	C/T	−0.083	0.019	1.72×10^{-5}	4.46×10^{3}	0.182	rs6073952	A/G	−0.082	0.016	2.53×10^{-7}	2.26×10^{5}	0.213	*
L5	*LIPC*	rs1532085	15	56470658	A/G	0.072	0.011	5.27×10^{-11}	3.04×10^{8}	0.351	rs4775041	C/G	0.061	0.010	1.41×10^{-9}	2.20×10^{6}	0.307	*
L7	*LIPC*	rs1532085	15	56470658	A/G	0.057	0.009	7.28×10^{-10}	2.20×10^{7}	0.351	rs4775041	C/G	0.041	0.008	2.26×10^{-7}	1.37×10^{4}	0.307	*
L4	*FADS1-2-3*	rs174546	11	61326406	T/C	−0.025	0.006	1.38×10^{-5}	1.59×10^{3}	0.303	rs102275	G/A	−0.025	0.006	1.41×10^{-5}	3.40×10^{2}	0.324	*
L6	*SORT1*	rs629301	1	109619829	C/A	−0.032	0.007	1.46×10^{-5}	—	0.219	—	—	−0.019	0.007	9.46×10^{-3}	—	0.219	
L1	*PLTP*	rs6065906	20	43987422	C/T	0.040	0.008	4.86×10^{-7}	1.58×10^{5}	0.182	rs6073952	A/G	0.034	0.007	7.79×10^{-7}	7.33×10^{4}	0.213	*
L8	*GCKR*	rs1260326	2	27584444	T/C	0.055	0.012	9.25×10^{-6}	—	0.425	—	—	0.035	0.011	1.45×10^{-3}	—	0.404	*
L8	*APOB*	rs1042034	2	21078786	G/A	−0.064	0.015	1.08×10^{-5}	3.21×10^{3}	0.237	—	—	−0.035	0.014	1.05×10^{-2}	2.38×10^{1}	0.196	
L8	*APOA1*	rs964184	11	116154127	G/C	0.126	0.018	4.82×10^{-12}	3.42×10^{3}	0.141	rs12286037	T/C	0.051	0.023	3.02×10^{-2}	4.92	0.061	
L10	*GCKR*	rs1260326	2	27584444	T/C	0.066	0.014	3.73×10^{-6}	—	0.425	—	—	0.049	0.013	2.91×10^{-4}	—	0.404	*
L10	*APOB*	rs1042034	2	21078786	G/A	−0.072	0.017	1.63×10^{-5}	2.13×10^{3}	0.237	—	—	−0.055	0.016	4.14×10^{-4}	6.02×10^{2}	0.196	*
L10	*APOA1*	rs964184	11	116154127	G/C	0.135	0.021	9.47×10^{-11}	1.74×10^{2}	0.141	rs12286037	T/C	0.049	0.030	1.11×10^{-1}	1.34	0.061	
L11	*APOA1*	rs964184	11	116154127	G/C	0.131	0.026	6.25×10^{-7}	—	0.141	rs12286037	T/C	0.071	0.037	5.53×10^{-2}	—	0.061	
L14	*GCKR*	rs1260326	2	27584444	T/C	0.095	0.022	2.01×10^{-5}	—	0.425	—	—	0.074	0.021	4.75×10^{-4}	—	0.404	*
L12	*GCKR*	rs1260326	2	27584444	T/C	0.081	0.018	6.88×10^{-6}	—	0.425	—	—	0.064	0.017	1.67×10^{-4}	—	0.404	*
L12	*APOA1*	rs964184	11	116154127	G/C	0.137	0.027	2.72×10^{-7}	—	0.141	rs12286037	T/C	0.070	0.038	6.54×10^{-2}	—	0.061	
L13	*APOA1*	rs964184	11	116154127	G/C	0.173	0.034	3.16×10^{-7}	—	0.142	rs12286037	T/C	0.095	0.051	6.29×10^{-2}	—	0.061	

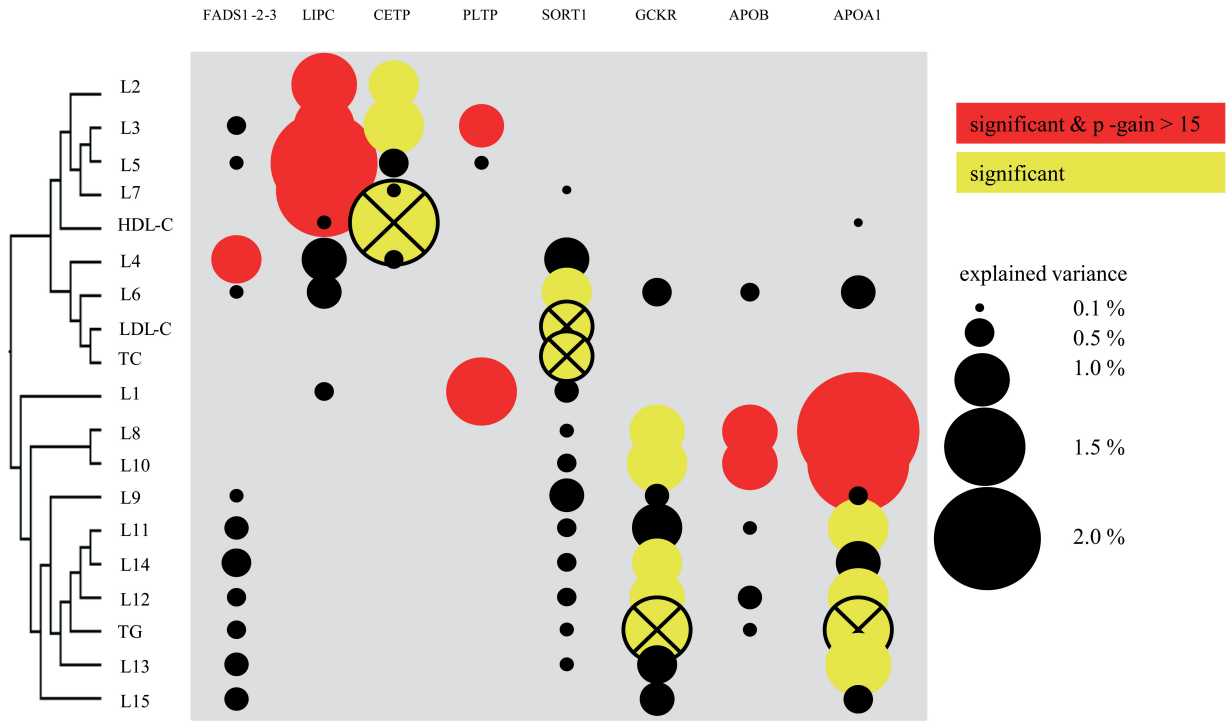

Figure 4.3: Explained variance of lipoprotein subfractions and serum lipids. This Figure presents the variance of the lipoprotein subfractions and serum lipids which is explained by the significantly associated loci. The explained variance is only shown for associations having a p-value < 0.05. The diameter of each circle represents the explained variance, a circle highlighted in yellow corresponds to a significant association and a circle coloured in red corresponds to a significant association with relevant p-gain. Circles with a black cross belong to serum lipids. The lipoprotein subfractions were ordered according to a hierarchical clustering which is displayed on the y-axis of this Figure (Petersen *et al.*, 2012).

When combining the observations made in the cluster analysis with the significant results of the association analysis, we detected comparable inter-relationships between the lipoprotein subfractions in both analyses. In the genetic analysis, we found that all lipoprotein subfractions of the cluster (L2, L3, L5, L7), which is correlated with HDL-C, were associated with *LIPC* whereas the subfractions L2 and L3 were also associated with *CETP*. With regard to subfraction L6 of the cluster (L4, L6) together with LDL-C and TC we found a significant association with *SORT1*. When considering the association between L4 and *SORT1*, we saw an effect although it was not significant (p-value $= 3.58 \times 10^{-5}$; Table 4.3). The subfractions L8 and L10, which built cluster (L8, L10), were associated with *GCKR*, *APOB* and *APOA1*. Subfractions L12 and L14 and subfractions L11, L12 and L13

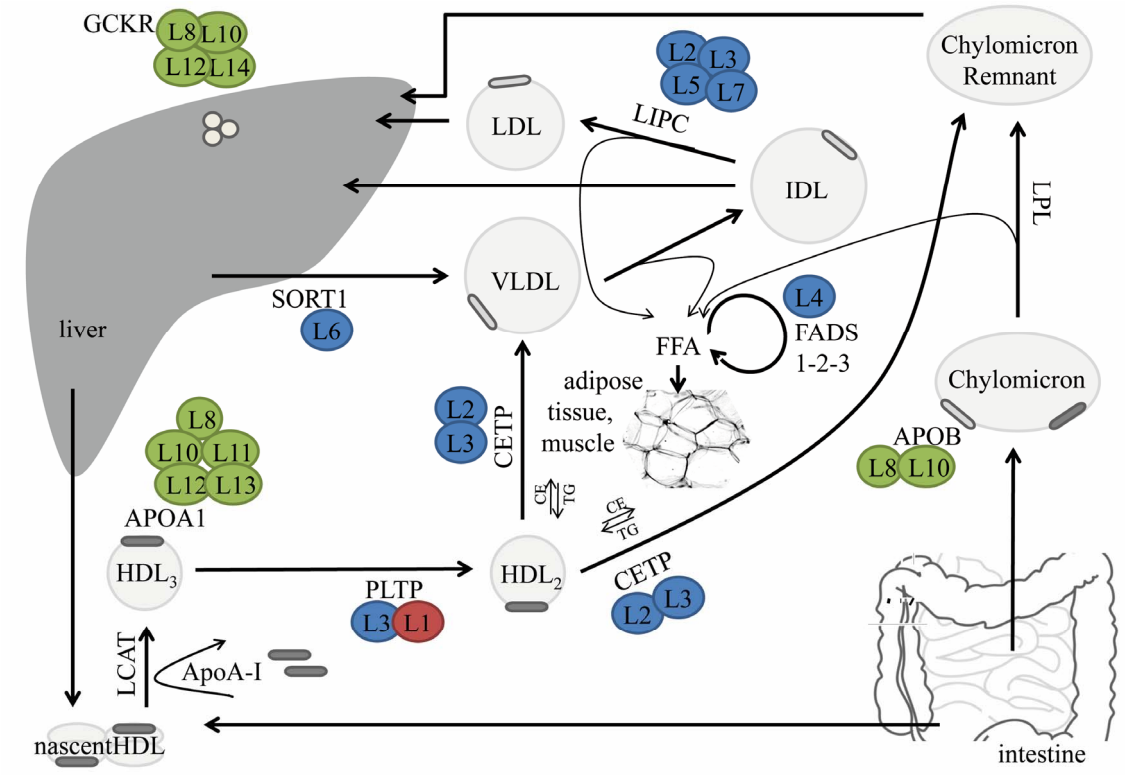

Figure 4.4: Classification of lipoprotein subfractions and their associated loci in the lipid metabolism. This Figure combines the results of the association analyses with the lipid metabolism. We displayed each associated gene at least once in this Figure and attached the associated lipoprotein subfractions to them. For clarity, we restricted the lipid metabolism to pathways where the associated loci are involved. The colour of the lipoprotein subfractions encodes the membership to a cluster. We assigned the lipoprotein subfractions to the three larger clusters L1, L2-L7 and L8-L15 to keep the Figure clear (Petersen *et al.*, 2012).

of cluster (L9, L11, L12, L13, L14, L15) together with TG were associated with *GCKR* and *APOA1*, respectively. Lipoprotein subfraction L1 was separate and only associated with *PLTP* with a relevant p-gain. Although lipoprotein subfraction L3 was also associated with *PLTP*, the effect was in opposite directions for L1 and L3 (Table 4.4). In conclusion, the genetic analysis confirms the observations made in the clustering and reveals further information about biological aspects of the lipoprotein subfractions.

Biological discussion

Clustering reveals that L1 is not captured by serum lipids

Clustering of the lipoprotein subfractions measured in fasting samples together with the serum lipids revealed five groups of subfractions. HDL-C clustered together with L2, L3, L5 and L7 whereas LDL-C and TC clustered together with L4 and L6 and TG clustered together with L9, L11, L12, L13, L14 and L15. In addition, we detected that lipoprotein subfraction L1 does not cluster together with the serum lipids. Due to its size, L1 is considered to correspond to the smallest HDL subfraction. This finding matches the observations made by others that the smallest HDL subfraction behaves in a different way than the larger HDL subfractions (Chasman *et al.*, 2009; Inouye *et al.*, 2010). Inouye *et al.* (2010) speculated that the smallest HDL subfraction may have pro-atherogenic potential which is in contrast to the anti-atherogenic properties of HDL-C. However, conflicting data on the association between cardiovascular disease risk and small HDL fractions still complicate painting a concise picture of the fractions' specific role (Camont *et al.*, 2011). HDL-C clustered together with L2, L3, L5 and L7 which are considered to correspond to medium and large HDL and very small and medium LDL, respectively. Interestingly, in addition to HDL related subfractions, LDL related subfractions also clustered together with HDL-C. Furthermore, LDL-C clustered together with L4 and L6 which are considered to be related to very large HDL and small LDL, respectively. This cross-mixed correlation of HDL and LDL subfractions needs further investigation. The subfractions clustered together with TG are related to the more TG-rich subfractions of VLDL and chylomicrons. When clustering the subfractions measured in plasma taken during a lipid tolerance test, we got different groups. The analysis of the lipoprotein subfractions during the lipid tolerance test revealed that some subfractions were increased on response to a standardised lipid tolerance test whereas other subfractions stayed nearly constant. While subfractions which cluster together with TG tend to increase after nutritional intervention, subfractions which cluster together with HDL-C stay the same. Interestingly, subfraction L13, which relates to remnants, behaves different than the other subfractions which cluster together with TG. Thus, nutritional intervention had different influences on distinct subfractions. The analysis of samples during the lipid tolerance test was carried out in only 15 subjects. However, HuMet is a highly controlled study and clustering of the subfractions at fasting time points led to a clustering comparable to that of KORA samples.

Using lipoprotein subfractions, we identified eight loci that were significantly associated in the KORA study whereas when analysing HDL-C, LDL-C, TC and TG in the same individuals we found only half of the loci. These eight loci contribute to diverse mechanisms of the lipid metabolism such as regulatory elements or structural lipid components which is illustrated in Figure 4.4.

PLTP indicates the role of L1 in the lipid metabolism

PLTP encodes for the phospholipid transfer protein which transfers phospholipids and other amphipathic compounds between lipoprotein particles (Huuskonen *et al.*, 2001; Rader, 2006) (Figure 4.4). Although the role of the phospholipid transfer protein in the reverse cholesterol transport has long been studied, it still remains controversial (Yazdanyar *et al.*, 2011). It has been shown in a large meta-analysis on serum lipids that *PLTP* is significantly associated with HDL-C and TG levels (Teslovich *et al.*, 2010) as well as with HDL particle size (Chasman *et al.*, 2009; Kaess *et al.*, 2011). Our analysis revealed that notably the lipoprotein subfraction L1, which was only weakly correlated with HDL-C, and lipoprotein subfraction L3 were associated with *PLTP* with opposite directions of effect. The other subfractions L2, L5 and L7 which clustered together with HDL-C showed no association. Here, the subfractions revealed an in-depth insight into the lipid metabolism. The opposite directions of effect of the association of L1 and L3 presumably compensate each other partly when analysing serum HDL-C. Moreover, due to the opposite directions of effect, it can be speculated that *PLTP* is involved in the conversion of L1 to L3 or vice versa. In addition, lipoprotein subfraction L1 was only marginally captured by the measurements of serum lipids as L1 was weakly negatively correlated with HDL-C and weakly positively with the other serum lipids. Therefore, it is possible that L1 is involved in parts of the lipid metabolism which were not covered by the measurement of the serum lipids. As L1 is related to the smallest HDL subfraction, it is assumed that L1 represents nascent HDL which would be an explanation for a negative correlation with HDL-C.

Lipoprotein subfractions revealed in-depth insight into mechanisms of *LIPC*, *CETP* and *FADS1-2-3*

LIPC encodes for hepatic lipase which catabolises TG-enriched HDL and breaksdown TG to diacyl- and monoacylglycerols and fatty acids (Rader, 2006). This molecular function is observed in associations between *LIPC* and numerous concentrations of glycerophosphatidylcholines, glycerophosphatidylethanolamines and sphingomyelins (Gieger *et al.*, 2008). In our analysis, the strongest association oc-

curred with L5 and L7 which clustered together with HDL-C and are considered to be related to very small and medium LDL, respectively. Here, we observed the largest increase in the proportion of explained variance compared to serum lipids. But also L2 and L3, the other lipoprotein subfractions which clustered together with HDL-C, were associated with *LIPC*. Interestingly, although all subfractions which cluster together with HDL-C were significantly associated with *LIPC* with the same direction of effect, the association between *LIPC* and HDL-C itself was not significant (p-value $= 1.60 \times 10^{-2}$, Table 4.3). For the remaining subfractions, especially for the subfractions correlated with TG, we did not see an association with *LIPC* as it is observed by others (Chasman *et al.*, 2009). Whereas *LIPC* was associated with all four lipoprotein subfractions which cluster together with HDL-C, *CETP* was only associated with L2 and L3. *CETP* encodes a protein which exchanges cholesteryl esters for TG between lipoproteins (Boes *et al.*, 2009) (Figure 4.4). The *FADS1-2-3* gene complex encodes for key enzymes in the metabolism of long-chain polyunsaturated fatty acids. Our analysis revealed an association between *FADS1-2-3* and L4, an LDL-C correlated subfraction which is considered to be related to large HDL. For LDL-C itself we did not see an association with *FADS1-2-3*. Although *FADS1-2-3* is strongly associated with TG in the global lipids meta-analysis in more than 100 000 samples (Teslovich *et al.*, 2010), we observed only a small effect which was not significant when based on the analysis of 1791 samples. The strong association between *FADS1-2-3* and L4 highlighted the potential of lipoprotein subfractions and hinted at further biological implications of the *FADS1-2-3* gene complex in the lipid metabolism.

More insight in pathway regulation and genes which encode structural components

Among others, *SORT1* and *GCKR* are genes that are involved in pathways regulating lipid and glucose metabolism. Musunuru *et al.* (2010) showed that hepatic expression of *SORT1* alters LDL-C and VLDL levels and that *SORT1* is associated with coronary artery disease. In more detail, *SORT1* encodes sortilin which presumably controls the biogenesis and hepatic release of VLDL from which LDL is generated by lipolysis (Kjolby *et al.*, 2010) (Figure 4.4). In our analysis, *SORT1* was associated with L6, which clustered together with LDL-C and relates to small LDL. *APOB* and *APOA1* are genes that encode the structural components apolipoprotein B and apolipoprotein A-I. Apolipoprotein B is the main apolipoprotein of chylomicrons, VLDL, IDL, LDL and lipoprotein(a) whereas apolipoprotein A-I is the main apolipoprotein of HDL (Kane *et al.*, 1980;

Rader, 2006) (Figure 4.4). In our analysis of KORA samples, both genes were predominantly associated with lipoprotein subfractions L8 and L10. These subfractions did not cluster closely with one of the serum lipids but were more related to the TG-correlated subfractions L9, L11, L12, L13, L14 and L15. These subfractions relate to VLDL as well as chylomicron subfractions. While *APOB* was only associated with L8 and L10, *APOA1* also showed associations with the particles L11, L12 and L13 in KORA. The associations of *APOA1* and *APOB* with L8 had the same direction of effect in KORA and GRAPHIC samples although we did not replicate them.

In total, we showed that lipoprotein subfractions provide a more detailed insight into the lipid metabolism and thus strengthen the association with disease-relevant genetic loci. Chasman *et al.* (2009) reported 43 loci associated with lipoprotein subfractions when analysing 17 296 women. At that time, ten of these loci were novel findings. By now, some of these loci were also found by Teslovich *et al.* (2010) in a serum lipid meta-analysis of more than 100 000 samples. Kaess *et al.* (2011) observed a strengthening in association when analysing HDL size and HDL particle number. In our results, we observed an increase in the proportion of variance explained when analysing lipoprotein subfractions instead of serum lipids. With the eight loci, we explained up to 4.5 % of the variance of the lipoprotein subfractions whereas only up to 3.3 % of the variance of serum lipids could be explained.

Overall, this study demonstrated that analysing well defined lipoprotein subfractions together with known genetic lipid loci leads to a genetic characterisation of the lipoprotein subfractions as well as an in-depth insight into various processes of the lipid metabolism. We identified five distinct groups of lipoprotein subfractions, one of them (L1) was only marginally captured by serum lipids and therefore extends our knowledge of lipoprotein biochemistry. During a lipid tolerance test, the relationship between the individual classes changed and L1 lost its special position. Based on this initial specification of the lipoprotein subfractions, further testing in clinical samples will reveal more information on their biological nature and their impact in disease causing mechanisms. All in all, NMR-based fine mapping of lipoprotein subfractions provides novel information on their biological nature and strengthens the association with genetic loci.

Conclusion

In this application of the candidate locus approach were lipoprotein subfractions analysed together with SNPs at 95 genetic lipid loci. This examination revealed an in-depth insight into biological pathways underlying the associations between the serum lipids and eight of the lipid loci. Moreover, the example of the *PLTP* locus showed that the analysis of lipoprotein subfractions together with candidate genes has the ability to detect opposed biological mechanisms which remained undetected in the analysis of serum lipids. In conclusion, this application confirmed and extended current knowledge about the lipid metabolism.

4.2 Application of the metabolomics GWAS approach: Human metabolic individuality in biomedical and pharmaceutical research

Background

Recent GWAS of metabolites have proven to be successful to reveal functional insight into biochemical mechansims (Gieger *et al.*, 2008; Hicks *et al.*, 2009; Tanaka *et al.*, 2009b; Chasman *et al.*, 2009; Illig *et al.*, 2010; Suhre *et al.*, 2011b; Demirkan *et al.*, 2012; Kettunen *et al.*, 2012). For instance, knowledge about the genetical basis of the β-oxidation or the biosynthesis of polyunsaturated fatty acids was gained (Gieger *et al.*, 2008; Illig *et al.*, 2010). Whilst in some of the metabolomics GWAS the analysis was focused on metabolite concentrations, others analysed also selected metabolite ratios or all pair-wise metabolite ratios. Despite the increased multiple testing burden when analysing all pair-wise metabolite ratios, this hypothesis-free approach brought promising results. For example, 14 out of 15 loci showed the strongest association with a metabolite ratio in Illig *et al.* (2010). However, one constraint of these metabolomics GWAS is that they were mostly based on lipid related metabolites. Extending the metabolomics GWAS approach to a broad set of metabolites covering many biochemical pathways will help to further understand the role of genetic predispositions for disease aetiology as well as to develop new and efficient therapies, among others.

The aim of this application is to gain more insight into the human metabolism through detection of novel genetic loci in an association analysis with over 250 blood metabolite concentrations as well as all pair-wise metabolite ratios. In addition to the GWAS, we link metabolic traits to clinically relevant phenotypes to gain further information about possible metabolic changes associated with biological processes underlying the clinically relevant phenotypes.

Results

In this application, we conducted GWAS of more than 250 metabolite concentrations as well as of about 37 000 pair-wise metabolite ratios in the KORA and TwinsUK studies using a step-wise approach. For the GWAS, we assumed an additive linear model and adjusted for age, sex and family structure. In most cases, this assumption was valid and there was no inflation of summary statistics

Table 4.5: Association data for significant SNPs in the meta-analysis. Number of samples (N), effect and other alleles (A/B), MAF, effect size (beta), p-value of association and p-gain are reported for the metabolic trait with the strongest association for KORA, TwinsUK and the meta-analysis. The loci are labeled by selected candidate genes (Suhre *et al.*, 2011a).

							KORA (N= 1768)		TwinsUK (N= 1052)		meta-analysis (N= 2820)		
locus	metabolic trait	SNP	chr	position	A/B	MAF	beta	p-value	beta	p-value	beta	p-value	p-gain
ACADS	butyrylcarnitine/propionylcarnitine	rs2066938	12	119644998	G/A	0.252	0.207	6.1×10^{-220}	0.203	5.1×10^{-79}	0.206	$< 4.4 \times 10^{-305}$	$> 10^{50}$
NAT8	N-acetylornithine	rs13391552	2	73672444	A/G	0.216	−0.213	8.9×10^{-149}	−0.181	8.5×10^{-67}	−0.201	5.4×10^{-252}	—
FADS1	1-arachidonoylglycerophospho-ethanolamine/1-linoleoylglycero-phosphoethanolamine	rs174547	11	61327359	C/T	0.320	−0.089	1.2×10^{-80}	−0.077	1.2×10^{-29}	−0.085	8.5×10^{-116}	5.7×10^{94}
UGT1A	bilirubin (E,E)/oleoylcarnitine	rs887829	2	234333309	T/C	0.337	0.112	1.2×10^{-56}	0.108	5.6×10^{-15}	0.112	2.9×10^{-74}	1.4×10^{42}
ACADM	hexanoylcarnitine/oleate (18:1n9)	rs211718	1	75879263	T/C	0.304	−0.080	4.4×10^{-53}	−0.066	2.4×10^{-17}	−0.076	2.2×10^{-71}	4.7×10^{15}
OPLAH	5-oxoproline	rs6558295	8	145211510	G/C	0.083	−0.061	8.4×10^{-51}	−0.039	4.7×10^{-9}	−0.056	1.5×10^{-59}	—
SCD	myristate (14:0)/myristoleate (14:1n5)	rs603424	10	102065469	A/G	0.189	0.051	5.3×10^{-43}	0.051	2.8×10^{-14}	0.051	2.9×10^{-57}	1.2×10^{48}
GCKR	glucose/mannose	rs780094	2	27594741	T/C	0.399	0.045	4.9×10^{-32}	0.038	6.1×10^{-21}	0.042	5.5×10^{-53}	1.5×10^{21}
NAT2	1-methylxanthine/ 4-acetamidobutanoate	rs1495743	8	18317580	G/C	0.188	−0.135	2.4×10^{-25}	−0.119	4.7×10^{-15}	−0.128	1.7×10^{-40}	1.1×10^{19}
CYP3A4	androsterone sulfate	rs17277546	7	99327507	A/G	0.045	−0.243	2.1×10^{-21}	−0.375	3.8×10^{-21}	−0.281	8.7×10^{-40}	—
ABO	ADpSGEGDFXAEGGGVR/ ADSGEGDFXAEGGGVR	rs612169	9	135133263	G/A	0.335	0.073	8.3×10^{-25}	0.098	1.6×10^{-16}	0.080	9.1×10^{-40}	4.1×10^{25}
SLC2A9	urate	rs4481233	4	9565177	T/C	0.193	−0.031	2.7×10^{-20}	−0.039	7.3×10^{-15}	−0.033	5.5×10^{-34}	—
CYP4A	10-nonadecenoate (19:1n9)/ 10-undecenoate (11:1n1)	rs9332998	1	47176773	C/T	0.135	−0.074	1.1×10^{-21}	−0.079	4.1×10^{-11}	−0.075	5.1×10^{-32}	2.5×10^{11}
CPS1	glycine	rs2216405	2	211325139	G/A	0.177	0.041	1.3×10^{-15}	0.069	1.9×10^{-14}	0.048	1.6×10^{-27}	—
LACTB	succinylcarnitine	rs2652822	15	61209825	C/T	0.467	−0.044	1.0×10^{-21}	−0.029	1.4×10^{-7}	−0.039	7.2×10^{-27}	—
SLC22A1	isobutyrylcarnitine	rs662138	6	160484466	G/C	0.163	−0.068	5.4×10^{-15}	−0.086	2.3×10^{-11}	−0.073	7.3×10^{-25}	—

Table 4.5 (cont.)

							KORA (N= 1768)		TwinsUK (N= 1052)		meta-analysis (N= 2820)		
locus	metabolic trait	SNP	chr	position	A/B	MAF	beta	p-value	beta	p-value	beta	p-value	p-gain
SLCO1B1	eicosenoate (20:1n9 or 11)/ tetradecanedioate	rs4149081	12	21269288	A/G	0.205	−0.098	2.0×10^{-13}	−0.109	3.7×10^{-9}	−0.102	2.8×10^{-22}	4.5×10^{8}
FUT2	ADpSGEGDFXAEGGGVR/ ADSGEGDFXAEGGGVR	rs503279	19	53900822	C/T	0.464	0.050	5.3×10^{-13}	0.061	8.5×10^{-8}	0.053	4.3×10^{-20}	2.9×10^{9}
ACE	aspartylphenylalanine	rs4329	17	58917190	G/A	0.465	−0.058	7.6×10^{-11}	−0.069	8.2×10^{-11}	−0.062	8.2×10^{-20}	—
PHGDH	serine	rs477992	1	120059099	A/G	0.313	−0.019	4.9×10^{-7}	−0.029	6.0×10^{-8}	−0.023	2.6×10^{-14}	—
ENPEP	ADpSGEGDFXAEGGGVR/ DSGEGDFXAEGGGVR	rs2087160	4	111554179	G/T	0.207	−0.048	1.1×10^{-7}	−0.093	3.6×10^{-7}	−0.057	6.5×10^{-13}	1.7×10^{7}
AKR1C	androsterone sulfate/ epiandrosterone sulfate	rs2518049	10	5128036	A/G	0.175	−0.027	4.8×10^{-6}	−0.039	1.1×10^{-7}	−0.032	6.7×10^{-13}	1.1×10^{10}
NT5E	inosine	rs494562	6	86173848	G/A	0.106	0.087	5.3×10^{-6}	0.189	1.2×10^{-9}	0.115	7.4×10^{-13}	—
PRODH	proline	rs2023634	22	17352450	G/A	0.091	0.054	4.3×10^{-21}	0.027	2.9×10^{-3}	0.046	2.0×10^{-22}	—
HPS5	alpha-hydroxyisovalerate	rs2403254	11	18281722	T/C	0.525	−0.053	2.6×10^{-16}	−0.039	1.4×10^{-5}	−0.048	1.0×10^{-20}	—
ALPL	ADpSGEGDFXAEGGGVR/ DSGEGDFXAEGGGVR	rs10799701	1	21693577	A/G	0.435	−0.060	2.2×10^{-15}	−0.062	1.9×10^{-5}	−0.061	2.9×10^{-20}	5.0×10^{14}
SLC7A6	glutaroyl carnitine/lysine	rs6499165	16	66883701	A/C	0.266	0.047	1.5×10^{-14}	0.041	4.1×10^{-5}	0.045	9.8×10^{-19}	1.4×10^{5}
KLKB1	bradykinin, des-arg(9)	rs4253252	4	187394452	T/G	0.492	−0.126	5.9×10^{-14}	−0.098	4.2×10^{-5}	−0.118	6.6×10^{-18}	—
GLS2	glutamine	rs2657879	12	55151605	G/A	0.187	−0.015	3.2×10^{-13}	−0.016	1.5×10^{-4}	−0.015	3.1×10^{-17}	—
PDXDC1	1-eicosatrienoylglycerophospho-choline/1-linoleoylglycerophospho-choline	rs7200543	16	15037471	G/A	0.304	−0.035	1.2×10^{-11}	−0.030	5.4×10^{-5}	−0.033	4.5×10^{-16}	5.9×10^{9}
SLC22A4	isovalerylcarnitine	rs272889	5	131693277	A/G	0.370	0.041	9.2×10^{-15}	0.021	1.1×10^{-2}	0.035	7.4×10^{-16}	—
AHR	caffeine/quinate	rs12670403	7	17275804	C/A	0.487	0.122	5.4×10^{-13}	0.083	4.0×10^{-3}	0.112	4.8×10^{-15}	2.3×10^{4}
ETFDH	decanoylcarnitine	rs8396	4	159850267	C/T	0.304	−0.050	2.3×10^{-12}	−0.034	4.7×10^{-4}	−0.045	5.5×10^{-15}	—
ELOVL2	docosahexaenoate (DHA; 22:6n3)/ eicosapentaenoate (EPA; 20:5n3)	rs9393903	6	11150895	A/G	0.242	−0.030	1.2×10^{-11}	−0.021	9.5×10^{-4}	−0.027	1.7×10^{-14}	6.7×10^{9}
SLC16A9	carnitine	rs7094971	10	61119570	G/A	0.147	−0.022	1.1×10^{-14}	−0.022	1.5×10^{-7}	−0.022	3.4×10^{-14}	—
IVD	3-(4-hydroxyphenyl)lactate/ isovalerylcarnitine	rs10518693	15	38487314	T/C	0.396	0.043	1.7×10^{-11}	0.028	2.9×10^{-3}	0.038	1.1×10^{-13}	1.3×10^{3}
SLC16A10	isoleucine/tyrosine	rs7760535	6	111853776	G/C	0.401	−0.017	2.1×10^{-10}	−0.012	4.5×10^{-3}	−0.015	1.4×10^{-12}	6.8×10^{5}

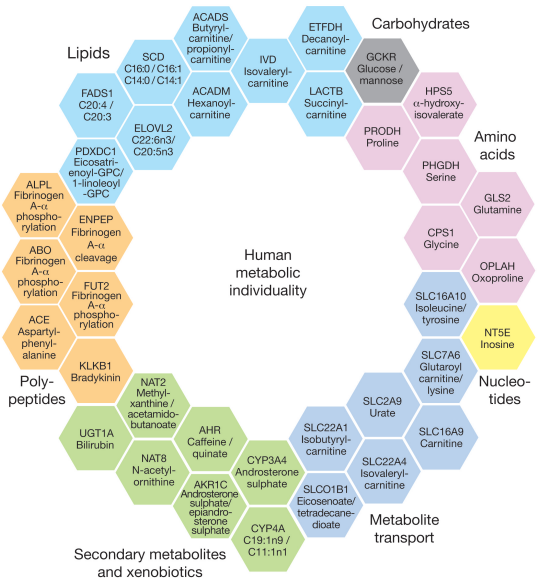

Figure 4.5: Thirtyseven loci associated with blood metabolites. This Figure summarises the 37 loci that were significantly associated with the analysed metabolic traits. Loci are shown colour coded by metabolic pathways together with selected associated metabolic traits (Suhre *et al.*, 2011a).

since λ values ranged from 0.940 to 1.024. After selection of promising genomic regions and metabolic traits in the first step of our GWAS, 666 SNPs and 643 metabolic traits remained. These SNPs and metabolic traits belonged to 115 independent signals. The regions and metabolic traits of these independent signals were further analysed using genotyped and imputed SNPs in KORA and TwinsUK, followed by a meta-analysis of both cohorts. This analysis revealed 37 loci which reached genome-wide significance after Bonferroni correction (Table 4.5 and Figure 4.5). Quantile-quantile plots for the GWAS of the metabolic traits which belong to the 37 loci are displayed in Figure A.1 in the Appendix. Since the observed distribution of the p-values coincided with the expected distribution of p-values for all except small p-values, we do not observe population stratification in our cohorts. The differences in levels of metabolic traits stratified by genotype are shown in the boxplots of Figure A.1. For metabolic traits such as butyrylcarnitine/propionylcarnitine and N-acetylornithine are the differences in the metabolite level apparent for the genoptypes of rs2066938 and rs13391552, respectively. In contrast, the stratification of decanoylcarnitine and isovalerylcarnitine by the genotypes of rs8396 and rs272889, respectively, revealed smaller but still significant variations. For 20 out of the 37 loci, the strongest association was observed

with a metabolite ratio. This observation confirms our strategy to conduct GWAS of both, metabolite concentrations and metabolite ratios.

For the selection of a candidate gene for each locus, we used the chromosomal position of the lead SNP to define a set of putative candidate genes (see Regional association plots of Figure A.1). Then, we used knowledge about the function of the genes as well as of the associated metabolic traits to determine one single candidate gene. The selected candidate genes are used to label the loci in Table 4.5. Afterwards, we used the catalogue of published GWAS (Hindorff *et al.*, 2009) to identify published associations between the 37 loci and clinically relevant phenotypes. For 15 loci, published associations could be identified (Table 4.6). Among others, associations with chronic kidney disease, metabolic syndrome, Crohn's disease and hypertriglyceridemia were identified as well as associations with risk factors for diseases such as serum lipids and fasting glucose-related traits.

Table 4.6: Published associations for genome-wide significant loci. This Table summarises the SNPs which are in LD$>$ 0.8 to the lead SNP and which have been reported to be associated with a clinically relevant phenotype in the catalogue of published GWAS (Hindorff *et al.*, 2011).

locus, SNP and metabolic trait	SNPs in LD that were reported in published GWAS; R^2 and D' to lead SNP	associated trait and reference
NAT8 rs13391552 N-acetylornithine	rs13538 $R^2 = 0.901$ D'$= 1.000$	chronic kidney disease (Köttgen *et al.*, 2010)
	rs10206899 $R^2 = 0.901$ D' $= 1.000$	serum creatinine (Chambers *et al.*, 2010)
FADS1 rs174547 1-arachidonoylglycerophosphoethanolamine/1-linoleoylglycerophosphoethanolamine	same SNP	resting heart rate (Eijgelsheim *et al.*, 2010) HDL-C (Kathiresan *et al.*, 2009)
	rs174550 $R^2 = 1.000$ D' $= 1.000$	fasting glucose-related traits (Dupuis *et al.*, 2010)
	rs174546 $R^2 = 1.000$ D' $= 1.000$	TC, HDL-C, TG (Teslovich *et al.*, 2010) LDL-C (Sabatti *et al.*, 2009; Teslovich *et al.*, 2010) metabolic syndrome (Zabaneh and Balding, 2010)

Table 4.6 (cont.)

locus, SNP and metabolic trait	SNPs in LD that were reported in published GWAS; R^2 and D' to lead SNP	associated trait and reference
	rs102275 $R^2 = 1.000$ D' = 1.000	Crohn's disease (Franke *et al.*, 2010)
	rs174583 $R^2 = 1.000$ D' = 1.000	response to statin therapy (Barber *et al.*, 2010)
	rs174601 $R^2 = 0.864$ D' = 0.963	alkaline phosphatase (Chambers *et al.*, 2011)
	rs174548 $R^2 = 0.800$ D' = 1.000	HDL-C, TG (Waterworth *et al.*, 2010)
UGT1A rs887829 bilirubin (EE)/oleoylcarnitine	same SNP	serum bilirubin levels (Sanna *et al.*, 2009; Chen *et al.*, 2012)
	rs6742078 $R^2 = 1.000$ D' = 1.000	serum bilirubin levels (Johnson *et al.*, 2009)
	rs4148325 $R^2 = 1.000$ D' = 1.000	bilirubin levels (Bielinski *et al.*, 2011)
GCKR rs780094 glucose/mannose	same SNP	fasting glucose-related traits, fasting insulin-related traits (Dupuis *et al.*, 2010) serum uric acid (Kolz *et al.*, 2009) TG (Willer *et al.*, 2008; Wallace *et al.*, 2008; Aulchenko *et al.*, 2009) C-reactive protein (Ridker *et al.*, 2008)
	rs780093 $R^2 = 1.000$ D' = 1.000	TG-Blood Pressure, Waist Circumference - TG (Kraja *et al.*, 2011) Crohn's disease (Franke *et al.*, 2010)
	rs1260326 $R^2 = 0.932$ D' = 1.000	platelet counts (Gieger *et al.*, 2011) gamma-glutamyl transferase (Chambers *et al.*, 2011) C-reactive protein (Dehghan *et al.*, 2011), TC (Teslovich *et al.*, 2010) TG (Kathiresan *et al.*, 2009; Teslovich *et al.*, 2010) hypertriglyceridemia (Johansen *et al.*, 2010) chronic kidney disease (Köttgen *et al.*, 2010) two-hour glucose challenge (Saxena *et al.*, 2010)

Table 4.6 (cont.)

locus, SNP and metabolic trait	SNPs in LD that were reported in published GWAS; R^2 and D' to lead SNP	associated trait and reference
	rs1260333 $R^2 = 0.870$ D' = 1.000	TG (Waterworth *et al.*, 2010)
NAT2 rs1495743 1-methylxanthine/4-acetamido-butanoate	rs1495741 $R^2 = 1.000$ D' = 1.000	bladder cancer (Rothman *et al.*, 2010) TC, TG (Teslovich *et al.*, 2010)
CYP3A4 rs17277546 androsterone sulfate	rs17277546 $R^2 = 1.000$ D' = 1.000	serum dehydroepiandrosterone sulphate levels (Zhai *et al.*, 2011)
ABO rs612169 ADpSGEGDFXAEGGGVR/ ADSGEGDFXAEGGGVR	rs514659 $R^2 = 1.000$ D' = 1.000	coronary heart disease (Reilly *et al.*, 2011)
	rs505922 $R^2 = 1.000$ D' = 1.000	venous thromboembolism (Trégouët *et al.*, 2009; Germain *et al.*, 2011) pancreatic cancer (Amundadottir *et al.*, 2009)
	rs657152 $R^2 = 0.931$ D' = 1.000	serum phytosterol levels (Teupser *et al.*, 2010) plasma levels of liver enzymes (Yuan *et al.*, 2008)
SLC2A9 rs4481233 urate	rs7442295 $R^2 = 0.871$ D' = 1.000	serum urate (Döring *et al.*, 2008; Wallace *et al.*, 2008)
SLC22A1 rs662138 isobutyrylcarnitine	rs1564348 $R^2 = 0.906$ D' = 1.000	TC, LDL-C (Teslovich *et al.*, 2010)
SLCO1B1 rs4149081 eicosenoate (20:1n9 or 11)/ tetradecanedioate	rs4363657 $R^2 = 1.000$ D' = 1.000	bilirubin levels (Bielinski *et al.*, 2011)
FUT2 rs503279 ADpSGEGDFXAEGGGVR/ ADSGEGDFXAEGGGVR	rs504963 $R^2 = 1.000$ D' = 1.000	Crohn's disease (McGovern *et al.*, 2010)
	rs281379 $R^2 = 0.966$ D' = 1.000	Crohn's disease (Franke *et al.*, 2010)
	rs602662 $R^2 = 0.933$ D' = 1.000	folate pathway vitamin levels (Tanaka *et al.*, 2009a)
	rs492602 $R^2 = 0.816$ D' = 1.000	TC (Teslovich *et al.*, 2010) plasma level of vitamin B12 (Hazra *et al.*, 2008)

Table 4.6 (cont.)

locus, SNP and metabolic trait	SNPs in LD that were reported in published GWAS; R^2 and D' to lead SNP	associated trait and reference
	rs516246 $R^2 = 0.816$ D' = 1.000	gamma-glutamyl transferase (Chambers *et al.*, 2011)
ACE rs4329 aspartylphenylalanine	rs4343 $R^2 = 0.816$ D' = 1.000	angiotensin-converting enzyme activity (Chung *et al.*, 2010)
ENPEP rs2087160 ADpSGEGDFXAEGGGVR/ DSGEGDFXAEGGGVR	rs6825911 $R^2 = 0.948$ D' = 1.000	blood pressure (Kato *et al.*, 2011)
ALPL rs10799701 ADpSGEGDFXAEGGGVR/ DSGEGDFXAEGGGVR	rs1780324 $R^2 = 1.000$ D' = 1.000	plasma levels of liver enzymes (Yuan *et al.*, 2008)
PDXDC1 rs7200543 1-eicosatrienoylglycerophosphocho-line/1-linoleoylglycerophosphocholine	rs1136001 $R^2 = 1.000$ D' = 1.000	height (Okada *et al.*, 2010)

After the identification of novel loci associated with blood metabolites, we further analysed them together with clinically relevant phenotypes. As first example, we selected from Table 4.6 the *NAT8* locus which is published to be associated with chronic kidney disease (Köttgen *et al.*, 2010). In our analysis, we observed an association between *NAT8* and N-acetylornithine. Therefore, we were interested whether N-acetylornithine was associated with eGFR which is a marker for kidney function. As a result, we found an association with eGFR in KORA and TwinsUK with p-value $= 7.6 \times 10^{-4}$ and p-value $= 3.6 \times 10^{-8}$, respectively, after adjusting for age and sex as well as family structure in TwinsUK.

Another approach to select clinically relevant phenotypes for the follow-up analysis is to use knowledge about gene function and biochemical pathways. An example where we applied this procedure is the *KLKB1* locus which encodes the kallikrein B plasma (Fletcher factor) 1. Plasma kallikrein is known to be involved in the regulation of blood pressure via the bradykinin pathway. This makes *KLKB1* a promising gene in a candidate gene analysis of hypertension (Lu *et al.*, 2007). Thus, we selected bradykinin which was associated with the *KLKB1* locus to investigate

the association with hypertension. As a result, this association analysis lead to a p-value $= 1.73 \times 10^{-9}$ and p-value $= 0.0495$ in KORA and TwinsUK, respectively, after adjustment for the covariates age and sex as well as family structure in the TwinsUK study.

Biological discussion

The discovered 37 loci may help to reveal further insight into biochemical mechanisms underlying the human metabolism. Therefore, we discuss exemplary the associations of *NAT8*, *KLKB1*, *ABO*, *FUT2*, *ALPL* and *ENPEP* in the following. Moreover, we take the examples of *FADS1* and *ACADS* to illustrate the strengthening in association when analysing metabolite ratios compared to raw metabolite concentrations.

NAT8 - N-acetylornithine - kidney function

An impaired kidney function is a risk factor for cardiovascular outcomes such as myocardial infarction and stroke. One measure to determine a reduced kidney function is the eGFR. Using this marker, GWAS have been conducted to investigate the genetical basis of kidney function (Chambers *et al.*, 2010; Köttgen *et al.*, 2010). Among others, the *NAT8* gene was identified in the 2p12-13 locus as a promising candidate. The *NAT8* gene encodes the N-acetyltransferase and is mainly expressed in the liver. Chambers *et al.* (2010) speculated that *NAT8* influences kidney function via the acetylation pathway which is an important mechanism for the detoxification process of medications as well as environmental toxins (Chambers *et al.*, 2010). Our metabolomics GWAS revealed an association between *NAT8* and N-acetylornithine which is involved in the acetylation process. Since the *NAT8* locus was already published to be associated with kidney function, we conducted a follow-up analysis and found an association between N-acetylornithine and eGFR. Therefore, our study confirmed the hypothesis that *NAT8* influences kidney function via the acetylation pathway. Nevertheless, causality cannot be inferred form our analysis and the clarification of the detailed processes needs further investigation. This was also pointed out by Nicholson *et al.* (2011) who found inconsistencies in the directionality of associations with the *NAT8* locus.

***KLKB1* - bradykinin - hypertension**

The second locus for which we conducted a follow-up analysis with a clinically relevant phenotype was *KLKB1*. In contrast to the *NAT8* locus, GWAS of hypertension as well as systolic and diastolic blood pressure did not reveal a significant association with *KLKB1* (Newton-Cheh *et al.*, 2009; Levy *et al.*, 2009; Ehret *et al.*, 2011). So far, these GWAS comprised of up to 200 000 samples and resulted in a p-value for rs4253252 in the *KLKB1* locus of 0.622 for systolic blood pressure and 0.221 for diastolic blood pressure (Ehret *et al.*, 2011). These p-values are far from being genome-wide significant. Nevertheless, *KLKB1* is a candidate gene for the analysis of hypertenstion (Lu *et al.*, 2007). *KLKB1* encodes the kallikrein B plasma (Fletcher factor) 1. Plasma kallikrein releases bradykinin in the blood and activates renin. Through these biochemical changes is blood pressure regulated by plasma kallikrein. Candidate gene studies showed an association between *KLKB1* and hypertension (Lu *et al.*, 2007). This context is supported by our study where we revealed an association between *KLKB1* and bradykinin which furthermore was associated with hypertension. One reason why it was not possible to detect the association between *KLKB1* and hypertension in GWAS, so far, might be that hypertension is influenced by many biochemical pathways. This pathway diversity is reflected in the broad spectrum of anti-hypertensive medications, e.g. angiotensin-converting-enzyme inhibitors, diuretics or beta blockers (Newton-Cheh *et al.*, 2009). However, it is essential to further investigate pathways involved in blood pressure regulation as well as to develop new anti-hypertensive drugs since a reduction of blood pressure achieves a reduction in risk for stroke, among others.

***GCKR* - mannose/glucose**

GCKR is localised on chromosome 2p23 and encodes the glucokinase (hexokinase 4) regulator (Warner *et al.*, 1995). Hitherto, it is known that this glucokinase regulating protein is oppositional influenced by fructose-6-phosphate and fructose-1-phosphate (Van Schaftingen, 1989; Malaisse *et al.*, 1990). Within the last years, several GWAS revealed that *GCKR* is a major pleiotropic risk locus. Associations with different clinically relevant phenotypes were reported, for example fasting glucose-related and fasting insulin-related traits (Dupuis *et al.*, 2010), serum uric acid (Kolz *et al.*, 2009), C-reactive protein (Ridker *et al.*, 2008) and serum lipids (Teslovich *et al.*, 2010) (Table 4.6). Our results showed an association of *GCKR* with the mannose/glucose ratio. This metabolite ratio was remarkably stronger associated with *GCKR* than the raw glucose concentrations (p-gain $= 1.5 \times 10^{21}$).

This finding may help to explain the observed associations between *GCKR* and some clinically relevant phenotypes as well as to further elucidate the role of mannose in the human metabolism. So far, it has been shown that mannose is used for the synthesis of glycoproteins (Taguchi *et al.*, 2005). For this glycosylation, mannose is formed from mannose-6-phosphate which can enter cells using a mannose specific transporter which is insensitive to glucose (Panneerselvam and Freeze, 1996; Taguchi *et al.*, 2005).

ABO, *FUT2* - ADpSGEGDFXAEGGGVR/ADSGEGDFXAEGGGVR and *ALPL*, *ENPEP* - ADpSGEGDFXAEGGGVR/DSGEGDFXAEGGGVR

All four loci (*ABO*, *FUT2*, *ALPL* and *ENPEP*) are associated in our study with a ratio of two fibrinogen A-α peptides. These peptides differ in the phosphorylation at serine. Additionally, the amino acid alanine is cleaved off in DSGEGDFXAEGGGVR compared to ADpSGEGDFXAEGGGVR for *ALPL* and *ENPEP*. An explanation for the association of fibrinogen ratios which represent fibrinogen phosphorylation with these loci might be through the phenotype alkaline phosphatase which is a liver enzyme that is used as a marker for biliary obstruction (Chambers *et al.*, 2011). The three loci *ABO*, *FUT2* and *ALPL* are known to be associated with alkaline phosphatase (Yuan *et al.*, 2008; Chambers *et al.*, 2011). Among others, the alkaline phosphatase is encoded by the *ALPL* locus. Furthermore, the association between the alkaline phosphatase and the *ABO* locus can be explained by an association between alkaline phosphatase and the ABO blood group (Whitfield and Martin, 1983). The *ABO* locus encodes a glycosyltransferase which is involved in the transfer of carbohydrates to the H antigen and thus encodes the ABO blood group antigens (Amundadottir *et al.*, 2009). Additionally, the expression of the ABO blood group antigens is also influenced by fucosyltransferase 2 which is encoded by *FUT2* (Hazra *et al.*, 2008). One may speculate now that the loci *ABO*, *FUT2* and *ALPL* influence the levels of alkaline phosphatase which furthermore may be linked to fibrinogen phosphorylation through a common pool of phosphate. The role of the *ENPEP* locus in this context is not clarified, so far. *ENPEP* encodes a glutamyl aminopeptidase (aminopeptidase A) and is known to be associated with blood pressure (Kato *et al.*, 2011).

Hitherto, GWAS of up to 22 000 samples have been conducted for blood fibrinogen concentrations. Despite these large sample sizes, none of them detected an association with any of the four loci (Dehghan *et al.*, 2009; Danik *et al.*, 2009; Lovely *et al.*, 2011). This observation may support the assumption that *ABO*, *FUT2*,

ALPL and *ENPEP* are associated with the phosphorylation of fibrinogen and not with raw fibrinogen concentrations.

***FADS1* - 1-arachidonoylglycerophosphoethanolamine/1-linoleoylglycerophosphoethanolamine**

The analysis of metabolite ratios strengthend the association compared to raw metabolite concentrations for the loci *GCKR*, *ABO*, *FUT2*, *ALPL* and *ENPEP*. Another example where this is the case is the *FADS1* locus. *FADS1* encodes the fatty acid desaturase 1 and was best associated with 1-arachidonoylglycerophosphoethanolamine/1-linoleoylglycerophosphoethanolamine in our study. The fatty acid desaturase which is encoded by *FADS1* is a key enzyme in the metabolism of long chain polyunsaturated omega 3 and omega 6 fatty acids where it converts dihomo-γ-linolenic acid (20:3n-6) to arachidonic acid (20:4n-6) (Lattka *et al.*, 2010). These metabolites have an association p-value of 1.03×10^{-4} for dihomo-linolenate (20:3n-3 or n-6) and of 2.3×10^{-21} for arachidonate (20:4n-6). For arachidonate (20:4n-6), the *FADS1* locus explains about 5.2 % of the observed variance. In contrast, the p-value for the association between the ratio of these metabolites, arachidonate (20:4n-6)/dihomo-linolenate (20:3n-3 or n-6) and the *FADS1* locus is 9.99×10^{-66} and the explained variance 15.3 %. This strengthening in association corresponds to the biological function of the *FADS1* gene. Thus, the biochemical properties of the associated metabolite pair provides information on the functional background of the association.

***ACADS* - butyrylcarnitine/propionylcarnitine**

Another example where the gene function matches the associated metabolic trait is the *ACADS* locus. The *ACADS* locus encodes an acyl-coenzyme A dehydrogenase which catalyses the β-oxidation of short chain acylcarnitines (Corydon *et al.*, 1997). In our study, the *ACADS* locus was associated with butyrylcarnitine/propionylcarnitine. Therefore, this ratio matches the substrate and product of the reaction of the short-chain acyl-coenzyme A dehydrogenase. Genes which belong to the same family as *ACADS* are *ACADM* and *ACADL*. *ACADM* encodes an enzyme which catalyses the β-oxidation of medium chain acylcarnitines whereas the enzyme encoded by *ACADL* catalyses the β-oxidation of long chain acylcarnitines. Associations between metabolic traits and the three acyl-coenzyme A encoding genes were observed by Illig *et al.* (2010).

Overall, this application revealed 37 loci that were associated with metabolic traits belonging to different biochemical classes. For two loci, we showed an association between a genetic variant, a metabolic trait and a clinically relevant phenotype: *NAT8* with N-acetylornithine and eGFR as well as *KLKB1* with bradykinin and hypertension. In total, the findings of this GWAS brought additional insight into pathways of the human metabolism and generated hypotheses to test in future studies.

Conclusion

This application of the metabolomics GWAS approach to more than 250 metabolite concentrations and over 37 000 pair-wise metabolite ratios revealed 37 loci to be involved in the human metabolism. Moreover, a follow-up analysis showed further associations between a metabolic trait and a clinically relevant phenotype for the two loci *NAT8* and *KLKB1*. All in all, this application confirmed and extended current knowledge about various processes of the human metabolism.

4.3 Statistical exploration of the p-gain: On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies

Background

The analysis of metabolite ratios has proven to be successful. As observed in Chapter 4.2, 20 out of 37 loci showed stronger associations with metabolite ratios than with metabolite concentrations. In order to quantify the strengthening in association when analysing metabolite ratios in comparison to metabolite concentrations, the p-gain was introduced (Gieger *et al.*, 2008). So far, the number of analysed metabolite concentrations was applied as an ad-hoc critical value of the p-gain. This approach can merely be regarded as an intuitive rule of thumb since a statistical determination of the distribution of the p-gain and herewith of the critical values has not yet been conducted.

Therefore, one aim of this thesis is to derive critical values through determination of the distribution of the p-gain and to provide a density table for readout of critical values. In addition, we investigate the characteristics of the p-gain in the situation of Bonferroni correction for multiple tests as well as the dependence of observed p-gain values on the sample size. Finally, we illustrate the power of the p-gain approach by investigating the enrichment for common pathways among metabolite ratios with large p-gain at the concrete example of the application of the metabolomics GWAS of Chapter 4.2.

Results and discussion

Formal definition of the p-gain

The p-gain was introduced in order to measure whether the association with a genetic locus is significantly stronger for a metabolite ratio than for the belonging metabolite concentrations. The definition of the p-gain for the ratio M_1/M_2 of metabolites M_1 and M_2 at a genetic locus X is as follows:

$$\text{p-gain}\left(\frac{M_1}{M_2}\middle|X\right) := \frac{\min(\text{p-value}(M_1|X), \text{p-value}(M_2|X))}{\text{p-value}(\frac{M_1}{M_2}|X)}. \qquad (1)$$

Conservative p-gain for common statistics

Although the p-gain was often used in metabolomics GWAS, only a rule of thumb for the determination of critical values was applied, so far. The p-gain was considered as being relevant when its value exceeded the number of analysed metabolite concentrations (Gieger *et al.*, 2008; Illig *et al.*, 2010, Chapter 4.1 and 4.2). Here, we derive critical values of the p-gain by determination of the distribution to define a more sensible threshold. As the distribution of the p-gain depends on the correlation structure among the metabolic traits, conservative critical values are beneficial in case of analysing multiple sets of metabolic traits since they can be applied to all analysed settings. For this purpose, we used an universal p-gain defined as the ratio of p-values belonging to two uncorrelated metabolic traits:

$$\text{p-gain}_{\text{univ}}\left(\frac{M_1}{M_2}\middle|X\right) := \frac{\text{p-value}(M_1|X)}{\text{p-value}(\frac{M_1}{M_2}|X)},\ \text{cor}(M_1, \frac{M_1}{M_2}) = 0. \tag{2}$$

Critical values of the distribution of this p-gain are conservative to the critical values of the distribution of the p-gain given in equation (1) (see Chapter 3.2.3). In the situation of the universal p-gain (equation (2)) we could use the convolution formula for density ratios which gave us a split density:

$$f_{\frac{P(M_1)}{P(M_1/M_2)}}(\text{p-gain}) = \begin{cases} \frac{1}{2\cdot\text{p-gain}^2}, & \text{p-gain} \geq 1 \\ \frac{1}{2}, & 0 < \text{p-gain} < 1 \end{cases},$$

which is displayed in Figure 4.6 (black line). To determine critical values, we derived the cumulative distribution function of the density, i.e.

$$F_{\frac{P(M_1)}{P(M_1/M_2)}}(\text{p-gain}) = \begin{cases} 1 - \frac{1}{2\cdot\text{p-gain}}, & \text{p-gain} \geq 1 \\ \frac{1}{2}\,\text{p-gain}, & 0 < \text{p-gain} < 1 \end{cases}.$$

Herewith, the critical value becomes $\frac{1}{2\cdot\alpha}$ with α denoting the level of significance. In the case of the typically used α level of 0.05, this yields a corresponding critical value for the p-gain of 10.

Critical values for multiple testing

In the case of conduction of many analyses, a correction for multiple testing has to be applied. When admitting a type I error rate of α and applying a Bonferroni

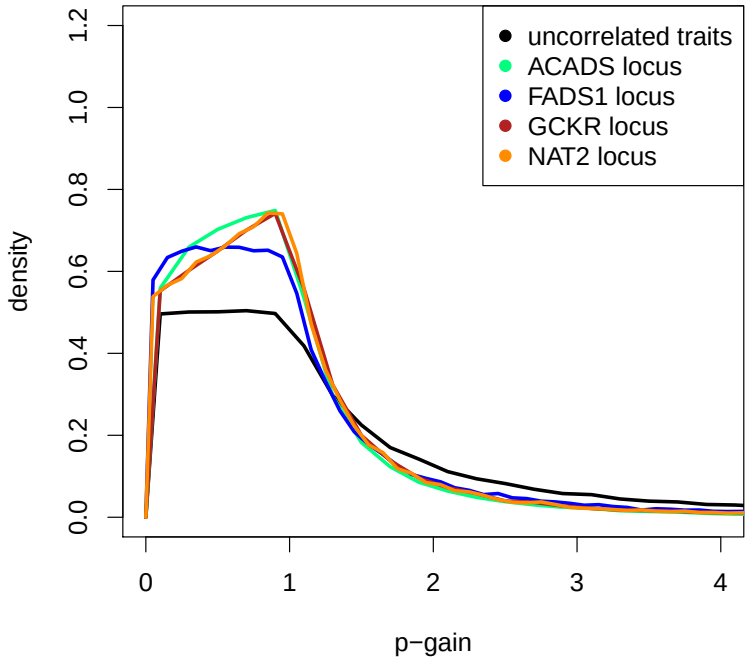

Figure 4.6: Density of the p-gain. This Figure shows the density of the p-gain for the calculated conservative p-gain of uncorrelated metabolic traits as well as for four loci which were significant in the application of the metabolomics GWAS approach (Chapter 4.2). The *ACADS* locus was found to be associated with butyrylcarnitine/propionylcarnitine, *FADS1* with 1-arachidonoylglycerophosphoethanolamine/1-linoleoylglycerophosphoethanolamine, *GCKR* with glucose/mannose and *NAT2* with 1-methylxanthine/4-acetamidobutanoate.

correction for B tests, i.e. aiming at a level of significance of $\frac{\alpha}{B}$, the critical value for the p-gain then becomes $\frac{B}{2 \cdot \alpha}$. For example, assumption of a type I error rate of $\alpha = 0.05$ leads to a critical value of $10 \cdot B$ which implies that for Bonferroni correction of B tests the uncorrected critical value of 10 can be multiplied by the number of tests B. Hence, the critical value of the p-gain in the situation of multiple testing is not the number of analysed metabolite concentrations, which was used so far as an ad-hoc criterion, but rather ten times the number of tests where the p-gain was applied.

P-gain for correlated metabolic traits

The case of uncorrelated metabolic traits (equation (2)) was conservative with respect to the p-gain as defined in equation (1). Therefore, we analysed also the density of the p-gain as defined in equation (1) for selected correlation settings.

In the situation of correlated metabolic traits the convolution formula cannot be applied anymore. Thus we simulated the density using a copula to generate the correlation among the metabolic traits. After this simulation, we transferred the correlation structure of the metabolic traits to a correlation structure among the p-values through conduction of linear regressions. Quantiles for the p-gain densities of correlated metabolic traits are provided in Table A.5 for various correlation settings. It can be observed that when any of the correlations $\text{cor}(M_1, M_1/M_2)$ or $\text{cor}(M_2, M_1/M_2)$ increase, the values of the quantiles of the p-gain decrease. This observation can be explained by the fact that the variation of the p-gain can be reduced by increasing the correlation between a metabolite concentration and the ratio (i.e. $\text{cor}(M_1, M_1/M_2)$ or $\text{cor}(M_2, M_1/M_2)$). A reduction of the variation of the p-gain leads to smaller critical values. On the other hand, for fixed $\text{cor}(M_1, M_1/M_2)$ and $\text{cor}(M_2, M_1/M_2)$, an increase in the correlation between M_1 and M_2 leads to an increase in the values for the p-gain quantiles when the correlation between M_1 and M_2 is not close to 0. Extending these observations to the most extreme and idealised case of having fully correlated metabolic traits which are uncorrelated with a third metabolic trait (i.e. $\text{cor}(M_1, M_2) = 1$, $\text{cor}(M_1, M_3) = 0$, $\text{cor}(M_2, M_3) = 0$) we get the largest critical values and thus these critical values are conservative to all correlation settings. Note that this idealised case is not possible for two metabolite concentrations M_1 and M_2 together with their ratio $M_3 = M_1/M_2$ as for fully correlated metabolite concentrations the ratio reduces to a numerical constant. This idealised case reduces the p-gain as defined in equation (1) to the universal p-gain as defined in equation (2). For this case, we derived the distribution using the convolution formula as well as through a simulation analysis. The results of both analyses coincided (Figure 4.7, Table A.5).

Dependence of p-gain values on sample size

In order to examine the behavior of the p-gain in the situation of real data, we computed the observed correlation structure among the metabolite ratios which were significant in the metabolomics GWAS of Chapter 4.2 (Table 4.7). This data set includes nearly uncorrelated metabolites, such as the ratio between 1-methylxanthine and 4-acetamidobutanoate (association with the *NAT2* locus) as well as highly correlated metabolites, such as the androsterone sulfate to epiandrosterone sulfate ratio (association with the *AKR1C* locus). The distributions of exemplary metabolite ratios are presented in Figure 4.6. As expected, the densities for correlated metabolic traits display smaller variations than the universal density for uncorrelated metabolic traits. Using this data set we conducted sim-

Table 4.7: Correlation among 20 significant metabolite ratios. This Table summarises the correlation structure among the 20 metabolite ratios which were discovered in the metabolomics GWAS of Chapter 4.2.

label	metabolite ratio (M_1/M_2)	correlation		
		(M_1;M_2)	(M_1;M_1/M_2)	(M_2;M_1/M_2)
ACADS	butyrylcarnitine/propionylcarnitine	0.422	0.769	−0.255
FADS1	1-arachidonoylglycerophosphoethanolamine/ 1-linoleoylglycerophosphoethanolamine	0.615	−0.547	0.323
UGT1A	bilirubin (E,E)/oleoylcarnitine	0.627	0.731	−0.073
ACADM	hexanoylcarnitine/oleate (18:1n9)	0.498	0.777	−0.159
SCD	myristate (14:0)/myristoleate (14:1n5)	0.830	−0.131	−0.662
GCKR	glucose/mannose	0.589	0.012	−0.801
NAT2	1-methylxanthine/4-acetamidobutanoate	0.038	0.896	−0.410
ABO	ADpSGEGDFXAEGGGVR/ADSGEGDFXAEGGGVR	0.407	0.724	−0.335
CYP4A	10-nonadecenoate (19:1n9)/10-undecenoate (11:1n1)	0.555	0.555	−0.383
SLCO1B1	eicosenoate (20:1n9 or 11)/tetradecanedioate	0.303	0.513	−0.662
FUT2	ADpSGEGDFXAEGGGVR/ADSGEGDFXAEGGGVR	0.407	0.724	−0.335
ENPEP	ADpSGEGDFXAEGGGVR/DSGEGDFXAEGGGVR	0.511	0.393	−0.589
AKR1C	androsterone sulfate/epiandrosterone sulfate	0.920	0.464	0.081
ALPL	ADpSGEGDFXAEGGGVR/DSGEGDFXAEGGGVR	0.511	0.393	−0.589
SLC7A6	glutaroyl carnitine/lysine	0.011	0.862	−0.497
PDXDC1	1-eicosatrienoylglycerophosphocholine/ 1-linoleoylglycerophosphocholine	0.579	0.676	−0.210
AHR	caffeine/quinate	0.207	0.748	−0.495
ELOVL2	docosahexaenoate (DHA; 22:6n3)/ eicosapentaenoate (EPA; 20:5n3)	0.771	0.203	−0.467
IVD	3-(4-hydroxyphenyl)lactate/isovalerylcarnitine	0.327	0.552	−0.607
SLC16A10	isoleucine/tyrosine	0.441	0.592	−0.462

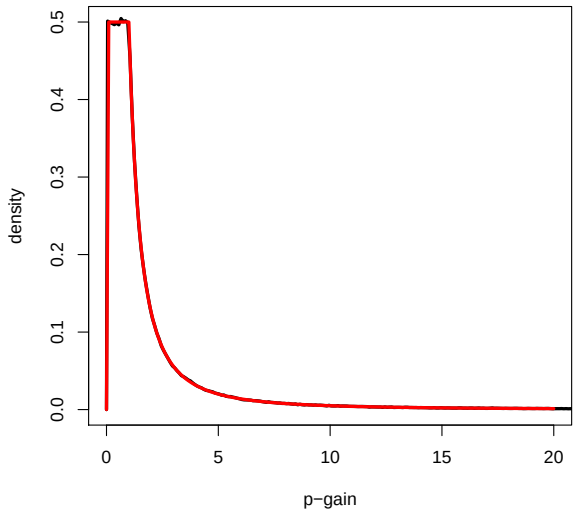

Figure 4.7: Calculated and simulated density for universal p-gain. This Figure shows that the calculated density of the universal p-gain and the simulated density of the p-gain for the idealised case of fully correlated metabolic traits which are uncorrelated with a third metabolic trait coincide. On the x-axis is the p-gain value entered and on the y-axis the density. The red line is the calculated density whereas the black line is the simulated density.

ulation tests to address the influence of the sample size on the observed p-gain values. We chose randomly sets of samples sizes between 100 and 2000 samples from the KORA study and calculated the p-gain for these sets. The results of this analysis illustrate the dependence of the p-gain values on the sample size (Table 4.8). For example, we observe for the association between the *ACADS* locus and the butyrylcarnitine to propionylcarnitine ratio a median p-gain value of 1.4×10^2 for a sample size of N= 100, of 1.1×10^5 for N= 500, of 2.8×10^{10} for N= 1000, of 3.1×10^{15} for N= 1500 and of 1.4×10^{21} for N=2000.

Pathway enrichment for metabolite ratios with large p-gains

To show the power of the p-gain approach, we conducted a pathway enrichment analysis for the 37 000 metabolite ratios of the application of the metabolomics GWAS (Chapter 3.2.2 and 4.2). Therefore, we compared the common pathway membership of metabolite ratios with a large p-gain in a GWAS to the overall average of common pathway affiliation in the metabolite ratio data set. Pathway membership of the metabolite concentrations was determined through evaluation of different pathway mappings (see Chapter 3.2.3). Moreover, we chose the largest p-gain of each metabolite ratio GWAS as allocation of a p-gain to each of the

Table 4.8: P-gain values for various sample sizes. This Table shows the dependence of the p-gain on the sample sizes for the 20 significant metabolite ratios of Chapter 4.2. The label of the locus, the metabolite ratio and SNP are given for each locus. Furthermore, the median as well as the 1st and 3rd quartiles are specified for randomly drawn sample subsets from the KORA study.

label	metabolite ratio	SNP	sample size				
			N= 100	N= 500	N= 1000	N= 1500	N=2000
ACADS	butyrylcarnitine/ propionylcarnitine	rs2066938	1.4×10^{2} (3.0×10^{0}; 6.4×10^{3})	1.1×10^{5} (8.3×10^{1}; 1.7×10^{11})	2.8×10^{10} (2.1×10^{3}; 4.9×10^{18})	3.1×10^{15} (8.6×10^{4}; 2.7×10^{27})	1.4×10^{21} (2.5×10^{7}; 2.0×10^{36})
FADS1	1-arachidonoyl-glycerophospho-ethanolamine/1-linoleoylglycero-phosphoethanol-amine	rs174547	1.4×10^{3} (1.2×10^{2}; 2.2×10^{4})	3.1×10^{8} (1.4×10^{3}; 4.1×10^{17})	4.1×10^{17} (2.2×10^{4}; 1.2×10^{32})	4.8×10^{25} (3.2×10^{8}; 3.3×10^{44})	2.7×10^{35} (1.6×10^{15}; 2.2×10^{58})
UGT1A	bilirubin (E,E)/ oleoylcarnitine	rs887829	4.1×10^{1} (7.3×10^{0}; 3.3×10^{2})	3.2×10^{4} (4.1×10^{1}; 5.6×10^{8})	4.9×10^{8} (3.2×10^{2}; 9.7×10^{14})	3.1×10^{12} (3.2×10^{4}; 4.3×10^{21})	2.1×10^{17} (1.1×10^{7}; 2.7×10^{28})
ACADM	hexanoylcarnitine/ oleate (18:1n9)	rs211718	3.7×10^{0} (7.4×10^{-1}; 1.7×10^{1})	1.8×10^{1} (1.5×10^{0}; 1.2×10^{3})	2.5×10^{2} (3.9×10^{0}; 2.0×10^{5})	6.1×10^{3} (1.0×10^{1}; 2.4×10^{7})	1.3×10^{5} (3.2×10^{1}; 3.4×10^{9})
SCD	myristate (14:0)/ myristoleate (14:1n5)	rs603424	5.1×10^{1} (9.5×10^{0}; 4.7×10^{2})	7.1×10^{4} (5.1×10^{1}; 4.1×10^{9})	3.5×10^{9} (4.6×10^{2}; 1.1×10^{17})	8.4×10^{13} (7.2×10^{4}; 9.1×10^{23})	2.7×10^{19} (8.2×10^{7}; 5.3×10^{31})
GCKR	glucose/mannose	rs780094	4.1×10^{0} (1.1×10^{0}; 2.7×10^{1})	1.6×10^{1} (1.8×10^{0}; 4.0×10^{2})	9.5×10^{1} (3.9×10^{0}; 1.8×10^{4})	7.8×10^{2} (8.4×10^{0}; 8.6×10^{5})	7.3×10^{3} (2.1×10^{1}; 2.6×10^{7})
NAT2	1-methylxanthine/ 4-acetamido-butanoate	rs1495743	1.7×10^{0} (6.1×10^{-1}; 4.2×10^{0})	7.6×10^{0} (1.4×10^{0}; 1.8×10^{2})	1.3×10^{2} (2.8×10^{0}; 1.5×10^{4})	2.3×10^{3} (7.2×10^{0}; 1.5×10^{6})	5.0×10^{4} (3.1×10^{1}; 1.5×10^{8})
ABO	ADpSGEGDFXA-EGGGVR/ADSG-EGDFXAEGGG-VR	rs612169	3.2×10^{0} (1.1×10^{0}; 1.0×10^{1})	4.2×10^{1} (3.0×10^{0}; 4.9×10^{3})	3.8×10^{3} (8.6×10^{0}; 5.6×10^{6})	3.2×10^{5} (4.1×10^{1}; 8.7×10^{9})	5.5×10^{7} (5.0×10^{2}; 1.1×10^{13})
CYP4A	10-nonadecenoate (19:1n9)/10-unde cenoate (11:1n1)	rs9332998	1.9×10^{0} (5.2×10^{-1}; 7.7×10^{0})	7.0×10^{0} (9.3×10^{-1}; 1.6×10^{2})	3.8×10^{1} (1.8×10^{0}; 4.6×10^{3})	2.9×10^{2} (3.5×10^{0}; 1.7×10^{5})	2.6×10^{3} (7.8×10^{0}; 6.0×10^{6})
SLCO1B1	eicosenoate (20:1n9 or 11)/ tetradecane-dioate	rs4149081	1.3×10^{0} (5.1×10^{-1}; 4.2×10^{0})	4.0×10^{0} (8.4×10^{-1}; 4.1×10^{1})	1.5×10^{1} (1.4×10^{0}; 4.6×10^{2})	6.8×10^{1} (2.6×10^{0}; 5.4×10^{3})	3.3×10^{2} (5.0×10^{0}; 6.3×10^{4})

Table 4.8 (cont.)

label	metabolite ratio	SNP	sample size				
			N= 100	N= 500	N= 1000	N= 1500	N=2000
FUT2	ADpSGEGDFXAEGGGVR/ADSGEGDFXAEGGGVR	rs503279	1.2×10^{0} (6.0×10^{-1}; 2.9×10^{0})	2.9×10^{0} (8.9×10^{-1}; 2.0×10^{1})	1.0×10^{1} (1.4×10^{0}; 2.0×10^{2})	4.4×10^{1} (2.3×10^{0}; 1.8×10^{3})	1.9×10^{2} (4.3×10^{0}; 1.8×10^{4})
ENPEP	ADpSGEGDFXAEGGGVR/DSGEGDFXAEGGGVR	rs2087160	1.0×10^{0} (4.6×10^{-1}; 2.4×10^{0})	1.9×10^{0} (7.0×10^{-1}; 8.8×10^{0})	4.3×10^{0} (9.4×10^{-1}; 3.8×10^{1})	1.0×10^{1} (1.4×10^{0}; 2.0×10^{2})	2.6×10^{1} (1.9×10^{0}; 9.1×10^{2})
AKR1C	androsterone sulfate/ epiandrosterone sulfate	rs2518049	1.3×10^{0} (5.5×10^{-1}; 4.1×10^{0})	3.0×10^{0} (8.8×10^{-1}; 1.9×10^{1})	8.3×10^{0} (1.3×10^{0}; 1.3×10^{2})	2.8×10^{1} (2.2×10^{0}; 8.1×10^{2})	8.7×10^{1} (3.4×10^{0}; 4.8×10^{3})
ALPL	ADpSGEGDFXAEGGGVR/DSGEGDFXAEGGGVR	rs10799701	1.5×10^{0} (5.6×10^{-1}; 4.5×10^{0})	4.6×10^{0} (9.2×10^{-1}; 5.4×10^{1})	2.4×10^{1} (1.7×10^{0}; 1.4×10^{3})	1.7×10^{2} (3.1×10^{0}; 3.6×10^{4})	1.3×10^{3} (6.7×10^{0}; 1.3×10^{6})
SLC7A6	glutaroyl carnitine/lysine	rs6499165	1.0×10^{0} (5.2×10^{-1}; 2.0×10^{0})	2.1×10^{0} (7.5×10^{-1}; 9.8×10^{0})	5.6×10^{0} (1.1×10^{0}; 6.9×10^{1})	1.8×10^{1} (1.7×10^{0}; 4.7×10^{2})	6.8×10^{1} (2.7×10^{0}; 3.5×10^{3})
PDXDC1	1-eicosatrienoyl-glycerophospho-choline/1-linoleoyl-glycerophospho-choline	rs7200543	1.3×10^{0} (5.7×10^{-1}; 4.0×10^{0})	2.8×10^{0} (7.6×10^{-1}; 1.8×10^{1})	6.8×10^{0} (1.1×10^{0}; 1.1×10^{2})	1.9×10^{1} (1.7×10^{0}; 7.1×10^{2})	5.9×10^{1} (2.7×10^{0}; 4.2×10^{3})
AHR	caffeine/quinate	rs12670403	1.0×10^{0} (4.6×10^{-1}; 2.6×10^{0})	2.3×10^{0} (6.5×10^{-1}; 1.7×10^{1})	5.7×10^{0} (9.0×10^{-1}; 1.0×10^{2})	1.8×10^{1} (1.3×10^{0}; 6.3×10^{2})	5.4×10^{1} (2.0×10^{0}; 5.1×10^{3})
ELOVL2	docosahexaenoate (DHA; 22:6n3)/ eicosapentaenoate (EPA; 20:5n3)	rs9393903	1.6×10^{0} (6.7×10^{-1}; 5.4×10^{0})	7.7×10^{0} (1.3×10^{0}; 1.4×10^{2})	7.6×10^{1} (2.9×10^{0}; 6.2×10^{3})	8.6×10^{2} (6.9×10^{0}; 3.3×10^{5})	1.1×10^{4} (2.0×10^{1}; 1.4×10^{7})
IVD	3-(4-hydroxy phenyl) lactate/ isovalerylcarnitine	rs10518693	1.0×10^{0} (5.3×10^{-1}; 2.3×10^{0})	2.1×10^{0} (7.1×10^{-1}; 1.3×10^{1})	5.5×10^{0} (9.5×10^{-1}; 7.4×10^{1})	1.6×10^{1} (1.3×10^{0}; 5.6×10^{2})	4.9×10^{1} (2.1×10^{0}; 3.1×10^{3})
SLC16A10	isoleucine/tyrosine	rs7760535	1.2×10^{0} (5.8×10^{-1}; 2.8×10^{0})	3.3×10^{0} (9.3×10^{-1}; 2.7×10^{1})	1.5×10^{1} (1.5×10^{0}; 2.9×10^{2})	6.9×10^{1} (2.8×10^{0}; 4.5×10^{3})	3.9×10^{2} (6.0×10^{0}; 7.4×10^{4})

37 000 metabolite ratios. Hence, we got a set of 37 000 'metabolite ratio - p-gain - SNP' assignments which we further analysed. As result, the observed p-gains varied from 10.02 to 1.68×10^{66} with a fast decrease in the highest values. Ascertainment of the metabolite ratios to pathway mapping revealed that on average 13.97 % of all metabolite ratios were on a pathway. In contrast, among the ten metabolite ratios with largest p-gain 57 % were mapped to a pathway (Table 4.9). For example, SNPs in the *FADS1* gene (rs174547) were associated with the ratio 1-

Table 4.9: Ten metabolite ratios with largest p-gain values. This Table summarises the ten metabolite ratios which have the largest p-gain values in our pathway enrichment analysis of 1768 KORA samples. For each metabolite ratio the associated SNP together with the effect size (beta), standard error (SE), p-value and p-gain of the association are provided. The pathway score specifies the percentage of pathway mappings which allocate both metabolites of the ratio to a common pathway.

metabolite ratio	SNP	gene	beta	SE	p-value	p-gain	pathway score (%)
1-arachidonoylglycerophosphoethanolamine/ 1-linoleoylglycerophosphoethanolamine	rs174547	*FADS1*	-0.09	0.004	1.15×10^{-80}	1.68×10^{66}	80
butyrylcarnitine/propionylcarnitine	rs2066938	*ACADS*	0.21	0.006	6.07×10^{-220}	6.15×10^{42}	80
myristate (14:0)/myristoleate (14:1n5)	rs603424	*SCD*	0.05	0.004	5.29×10^{-43}	4.22×10^{35}	80
bilirubin (E,E)/oleoylcarnitine	rs887829	*UGT1A*	0.11	0.007	1.15×10^{-56}	2.63×10^{32}	33
1-arachidonoylglycerophosphocholine/ 1-eicosadienoylglycerophosphocholine	rs174577	*FADS2*	-0.09	0.006	7.79×10^{-52}	1.34×10^{32}	60
ADpSGEGDFXAEGGGVR/ADSGEGDFXAEGGGVR	rs612169	*ABO*	0.07	0.007	8.27×10^{-25}	5.06×10^{15}	80
myo-inositol/N-acetylornithine	rs7607014	*ALMS1*	0.21	0.007	6.35×10^{-158}	1.05×10^{15}	17
butyrylcarnitine/palmitate (16:0)	rs10431384	*MLEC*	0.17	0.006	3.38×10^{-136}	1.58×10^{13}	20
hexanoylcarnitine/oleate (18:1n9)	rs12134854	*ACADM*	-0.08	0.005	9.17×10^{-54}	5.42×10^{10}	20
acetylcarnitine/hexanoylcarnitine	rs6699682	*MSH4*	0.07	0.006	7.146×10^{-40}	1.21×10^{10}	100

arachidonoylglycerophosphoethanolamine/1-linoleoylglycerophosphoethanolamine, among others. This metabolite ratio was mapped to the metabolic pathways of biosynthesis of unsaturated fatty acids and the linoleic acid metabolism (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2006, 2010). It has been shown that the delta-5 desaturase, which is encoded by the *FADS1* gene, converts dihomo-γ-linolenic acid (20:3n-6) to arachidonic acid (20:4n-6) and eicosatetraenoic acid (20:4n-3) to eicosapentaenoic acid (20:5n-3) (Lattka *et al.*, 2010). Therefore, the metabolite ratios which were associated in our analysis with SNPs in the *FADS1* gene match the known function of the delta-5 desaturase. Another example is the *ACADS* locus. Here, we observed an association with the metabolite ratios butyrylcarnitine/propionylcarnitine, among others. Metabolites of this ratio are quaternary amines and were mapped to the pathway of carnitine metabolism (Evans *et al.*, 2009). The *ACADS* locus encodes a gene of the acyl-coenzyme A dehydrogenase family. This enzyme catalyses the initial step of the mitochondrial fatty acid β-oxidation pathway. Among others, increased butyrylcarnitine, or 'C4 carnitine', is a biomarker for short chain acyl-coenzyme A dehydrogenase deficiency (Jethva *et al.*, 2008). In addition to these results of the ten ratios with largest p-gain, among the metabolite ratios with the 100 largest p-gains 49.10 % were mapped to common pathways. The difference to the overall average of 13.97 % corresponds to a p-value of 7.24×10^{-17}. When examining the metabolite ratios with the largest 500, 1000 and 1500 p-gains, still 34.90 %, 29.13 % and 25.66 %, respectively, were on the same pathway. The entire development of pathway allocation of metabolite ratios is displayed in Figure 4.8. Moreover, among the metabolite ratios with significant p-gain after Bonferroni correction, i.e. p-gain $\geq 10 \cdot 37\,000$, 43.57 % were on a common pathway compared to 13.8 % for metabolite ratios with a p-gain $< 10 \cdot 37\,000$. This difference corresponds to a p-value of 9.64×10^{-26}. These results highlight the impact of metabolite ratios together with the p-gain as a useful tool when analysing 'omics data.

Conclusion

Taken together, we showed that the p-gain is an appropriate measure for large scale 'omics data which emphasises metabolite ratios enriched for biochemical pathways. For the p-gain, we derived critical values to determine significance for various situations. Given the success of the approach in the metabolomics field, hypothesis free testing of ratios between biologically related quantitative traits should also be considered for association studies with other 'omics data sets.

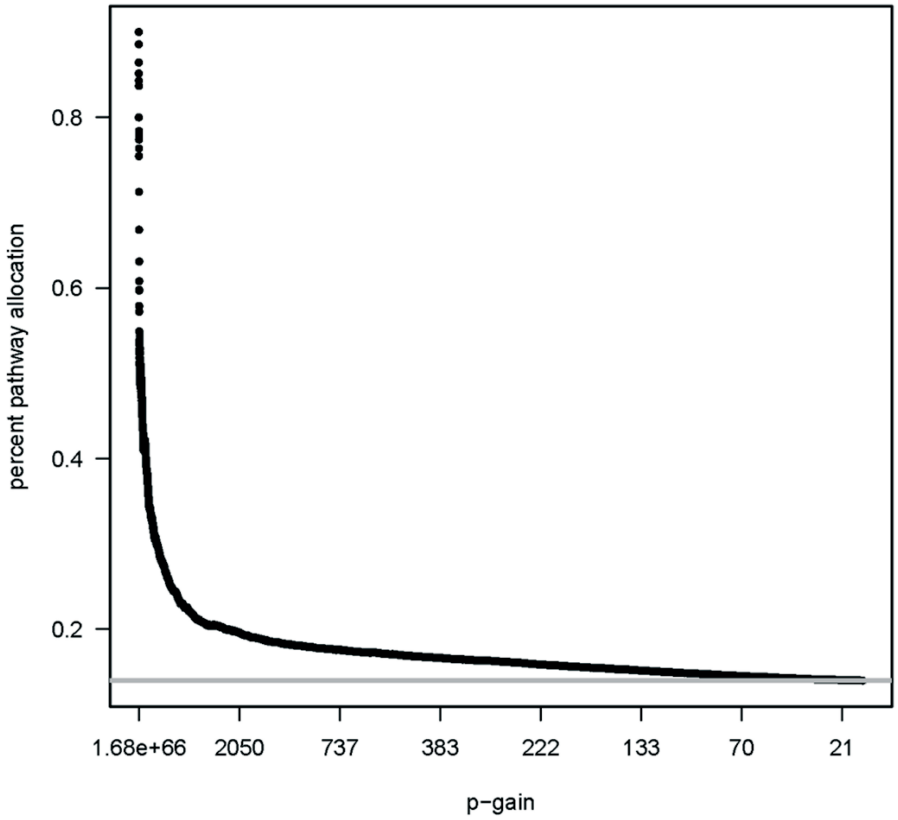

Figure 4.8: Mean pathway membership among metabolite ratios across different p-gain sizes. This Figure depicts the relationship between the p-gain and the pathway allocation for metabolite ratios. The underlying data set is composed of the SNP association of each metabolite ratio GWAS which yielded to the largest p-gain. The x-axis carries the p-gain whereas the percent of pathway allocation is entered on the y-axis. The grey line represents the overall average of 13.97 % pathway allocation. The black line represents the cumulative mean of the pathway allocation, beginning with the metabolite ratios with the largest p-gains, i.e. the first point corresponds to the pathway allocation of the metabolite ratio with largest p-gain, the second to the mean pathway allocation for the two metabolite ratios with largest p-gains, the third to the mean pathway allocation for the three metabolite ratios with largest p-gains, ... and the last point corresponds to the mean pathway allocation across all analysed metabolite ratios and therefore coincides with the grey line of overall percentage of pathway allocation.

5. Discussion and Conclusion

In this thesis two procedures for the integration of metabolomics data in the GWAS approach are presented and applied to concrete examples. The candidate locus approach utilises metabolites in order to gain further understanding of the processes underlying known associations between genetic loci and clinically relevant phenotypes. This approach was applied to a data set of 15 lipoprotein subfractions to reveal novel information about the role of 95 known lipid loci in the lipid metabolism. As a result, significant associations with lipoprotein subfractions were detected for eight of the analysed loci. Additionally, for five of these loci a strengthening in association was observed when analysing lipoprotein subfractions compared to serum lipids.
In the metabolomics GWAS approach, hypothesis-free analysis of metabolic traits are conducted to discover novel genetic loci which serve as candidate loci for further analyses with clinically relevant phenotypes. This approach was applied to a metabolomics data set comprising of more than 250 metabolites covering 60 different pathways and all pair-wise metabolite ratios with the aim to find novel loci associated with blood metabolites. These GWAS resulted in the discovery of 37 loci which belonged to different metabolic pathways. In a follow-up analysis, the detected associations between *KLKB1* and bradykinin as well as *NAT8* and N-acetylornithine were further investigated with respect to the phenotypes hypertension and eGFR, respectively.
Finally, the p-gain concept is statistically explored in this thesis. In detail, the distribution of the p-gain and its critical values were derived for different correlation settings among the metabolic traits and the power of the p-gain approach was shown in a pathway enrichment analysis. This statistical exploration of the p-gain improved the analysis of metabolite ratios substantially.

The two presented procedures incorporate metabolites in the GWAS approach in different ways. Despite their different proceeding, the objective of both approaches is to gain knowledge about genetical and biochemical mechanisms underlying clin-

ically relevant phenotypes. Thus, an improved understanding of metabolic processes can lead to the specification of new biomarkers for disease detection and prediction, to the development of new drug targets or the elucidation of adverse reactions to medication. Regarding this aim of an improved understanding of biochemical mechanisms, the application of the candidate locus approach revealed further insight into the role of the *PLTP* locus in the HDL metabolism, among others. So far, the behaviour of HDL in atherosclerosis is not completely clarified. To this end, the detected associations of L1 and L3 with *PLTP* may help to resolve some of the ambiguities of HDL. In connection with the application of the metabolomics GWAS approach, knowledge was gained about different biochemical mechanisms. For example, associations were discovered which yielded insight into the bradykinin pathway or nephrotic detoxification processes. It is essential to understand these pathways as the bradykinin pathway is involved in the regulation of blood pressure whereas a reduced ability to detoxify nephrotic medications can lead to impaired kidney function.

Despite the comparable aim of both procedures, each procedure has its own advantages and limitations. A characteristic of the candidate locus approach is that only genetic loci which were already detected in GWAS of clinically relevant phenotypes are further analysed. This restriction to preselected candidate loci is an advantage since it results in a reduced multiple testing burden compared to the metabolomics GWAS approach. For example, in the application of the candidate locus approach, 101 SNPs at 95 lipid loci were analysed together with 15 lipoprotein subfractions leading to a Bonferroni corrected level of significance of 3.3×10^{-5}. In contrast, more than 250 metabolite concentrations and about 37 000 pair-wise metabolite ratios were analysed on 600 000 genome-wide SNPs in the application of the metabolomics GWAS approach. This resulted in a Bonferroni corrected level of significance of 2.0×10^{-12}.
Another advantage of the candidate locus approach is that existing knowledge about relationships between genes and phenotypes is applied and extended. This knowledge was gained in GWAS which comprised of tens of thousands of samples. For instance, the 95 lipid loci analysed in the application of the candidate locus approach were discovered in GWAS of more than 100 000 samples of 46 different studies (Teslovich *et al.*, 2010).
Despite these advantages, the restriction to candidate loci is also a limitation of this approach. Metabolites are more refined phenotypes than most other clinically relevant phenotypes which often represent aggregated variables comprising of dif-

ferent sub-phenotypes. For example, four HDL related lipoprotein subfractions were analysed in the application of the candidate locus approach instead of aggregated HDL-C. As a consequence, with the utilisation of metabolites it is possible to discover loci which have not been detected in GWAS of the aggregated phenotypes. With the detection of additional loci, metabolites can help to elucidate parts of the missing heritability of the related aggregated phenotypes. Due to the restriction to candidate loci, this procedure does not have the ability to discover novel loci. Nevertheless, the example of the *PLTP* locus illustrates this ability of metabolites. Regarding the minor allele of rs6065906, *PLTP* has an increasing effect on L1 and a decreasing effect on L3. In the analysis of aggregated HDL-C these opposite effects cancel out each other partly leading to a small decreasing effect of the minor allele of rs6065906 on HDL-C. To discover loci with a small effect size, large sample sizes are necessary as it was the case in Teslovich *et al.* (2010).

In contrast to the candidate locus approach, the possibility to discover novel loci is a strength of the metabolomics GWAS approach. In total, 37 loci were detected when analysing about 2800 samples in the application presented in this thesis. For many of these loci, the function of the gene matches the associated metabolic trait. One example is the *ACADS* locus which is associated with butyrylcarnitine/propionylcarnitine. These metabolites are a substrate-product pair of acyl-coenzyme A dehydrogenases which are encoded by *ACADS*. Another example is the *NAT8* locus which is associated with N-acetylornithine. The *NAT8* locus encodes N-acetyltransferase whose function matches N-acetylornitine. This illustrates that the biological mechanisms underlying an association are easier to understand for associations with metabolic traits than for associations with clinically relevant phenotypes.

A limitation of the second procedure is the multiple testing burden as already described. The level of significance applied in the application of the metabolomics GWAS approach was set to 2.0×10^{-12}. This level was derived by Bonferroni correction for all tested pair-wise metabolite ratios as well as for all tested genome-wide SNPs. In the situation of metabolomics GWAS, Bonferroni correction is very conservative since many SNPs are in LD and some of the metabolites are highly correlated due to a close biological relationship. Moreover, the amount of correlated metabolites is artificially increased in case of analysing all pair-wise metabolite ratios.

Another limitation of the metabolomics GWAS approach is the computational as well as data storage burden, especially if all pair-wise metabolite ratios were

analysed. Whilst it is possible to carry out an application of the candidate locus approach using an usual personal computer, it is necessary to have a large linux cluster as well as appropriate data storage devices available for the feasibility of the metabolomics GWAS.

Strategies which are similar to the two procedures presented in this work were also applied by others to incorporate metabolomics data in the GWAS approach. For example, a study published by Tukiainen *et al.* (2012) uses lipoprotein subfractions and lipid related metabolites to further characterise the 95 lipid loci published by Teslovich *et al.* (2010). This study is comparable to the application of the candidate locus approach presented in this thesis as it also analyses lipoprotein subfractions together with known genetic lipid loci. Despite this, the objectives of both studies are different. The study by Tukiainen *et al.* (2012) further characterises the lipid loci through associated lipoprotein subfractions, aims at detecting causal variants through a fine-mapping approach of the loci and searches for independent genetic signals in the loci. In contrast, our application characterised the lipoprotein subfractions through a clustering with serum lipids and an analysis of samples during nutritional intervention followed by a mutual characterisation of the subfractions and lipid loci in a genetic association analysis. This lead to further insight into the lipid metabolism.
Concerning the second procedure, examples of metabolomics GWAS are the publications of Gieger *et al.* (2008), Hicks *et al.* (2009), Tanaka *et al.* (2009b), Chasman *et al.* (2009), Illig *et al.* (2010), Suhre *et al.* (2011b), Lemaitre *et al.* (2011), Kettunen *et al.* (2012) or Demirkan *et al.* (2012). Within these studies, different metabolites were investigated to gain a better understanding of the genetics underlying the analysed metabolites. The application of the metabolomics GWAS presented in this work is in-line with this approach. As an extension of the metabolomics GWAS, for some of the detected loci associations with additional clinically relevant phenotypes were examined. Another aspect to mention is that beside metabolite concentrations also all pair-wise metabolite ratios were analysed in the application presented here. This was also done by Illig *et al.* (2010) whereas others analysed only selected ratios, e.g. Hicks *et al.* (2009), Kettunen *et al.* (2012), or focused solely on the analysis of metabolite concentrations, e.g. Tanaka *et al.* (2009b).

So far, this hypothesis-free analysis of all possible metabolite ratios has proven to be successful even if it increases the multiple testing burden. Furthermore, the p-gain was used together with a provisional cut-off rule as an objective measure of the increase in information. To improve the application of the p-gain, the second aim of this thesis was to conduct a statistical exploration. As a result, the critical p-gain value after Bonferroni correction for B tests is at $\frac{B}{2 \cdot \alpha}$ with α being the nominal significance level. This finding implicates that the critical value of the p-gain in Chapter 4.1 should be corrected for the 21 significant SNP - lipoprotein subfraction associations. This leads to a critical p-gain value of $210 = 21 \cdot 10$ instead of 15. When we apply this critical value to the results, the association between *APOA1* and L10 has no significant p-gain value anymore. In addition to this, the results of the statistical p-gain exploration should also be applied to the results in Chapter 4.2. Here, the critical p-gain value of 10 should be corrected for the number of tests where the p-gain was applied. This number depends on the analysis strategy. In the case of a one step approach a simultaneous filter is applied to the p-value and p-gain. The number of tests is equal to the number of calculated associations between metabolite ratios and SNPs, which was approximately $37\,000 \cdot 600\,000$ in Chapter 4.2. In the situation of a step-wise approach where in a first step a p-value filter is applied and in a second step a filter for a p-gain, this number will be smaller. However, to consider a p-gain larger than 250 as relevant, as it was done in Chapter 4.2, is not accurate anymore.

As another consequence of the exploration of the p-gain, it is now possible to evaluate the GWAS of the metabolite ratios of Chapter 4.2 according to a significant p-gain instead of a significant p-value. Such an evaluation will reveal metabolite ratios which are significantly better associated with a genetic locus than the corresponding single metabolite concentrations. For this purpose, the SNP with the largest p-gain was determined for each metabolite ratio and these 37 000 'metabolite ratio - SNP - p-gain' sets were tested according to a p-gain larger than the critical p-gain value of $370\,000 = \frac{37\,000}{(2 \cdot 0.05)}$. As a result, some loci were detected that were not among the loci reported in Chapter 4.2, e.g. *MLEC* or *MSH4*. Since this evaluation of the GWAS of metabolite ratios according to a significant p-gain was only started in Chapter 4.3, further extinctive explorations are needed.

Finally, we showed a dependence of the observed p-gain values of the metabolomics GWAS on the sample size. Building on the knowledge gained about the distribution of the p-gain, it is now possible to conduct an analysis of the statistical power of the p-gain.

In total, both presented procedures have proven to be successful as they confirmed and extended current knowledge about different genetical and biochemical mechanisms. As a consequence of their distinct advantages and limitations, the procedures exploit different properties of the metabolomics data and thus complement each other. Hence, for a most extensive evaluation of metabolomics data it is preferable to utilise both procedures. Furthermore, it is recommendable to evaluate metabolite ratios together with the p-gain as an objective measure. Overall, this thesis proved that the incorporation of metabolites in the GWAS approach is a promising way to gain understanding of the genetical and biochemical mechanisms underlying disease aetiology. An expansion of the discussed procedures to the incorporation of multiple 'omics technologies such as transcriptomics, proteomics or epigenomics will lead towards a further understanding of complex diseases such as type 2 diabetes or cardiovascular diseases.

Appendix

Table A.1: Definition of lipoprotein subfractions L1-L15. The lipoprotein subfractions L1-L15 and their correspondence to subfractions defined by Linsel-Nitschke *et al.* (2009) (Petersen *et al.*, 2012).

NMR lipoprotein subfraction	related lipoprotein subfraction	particle diameter [nm]	average density [g/ml]
L1	small HDL	7−8.5	1.200
L2	medium HDL	8.5−10	1.120
L3	large HDL	10−13	1.090
L4	very large HDL	13−16	1.063
L5	very small LDL	16−19	1.060
L6	small LDL	19−21	1.045
L7	medium LDL	21−22	1.035
L8	large LDL	22−25	1.027
L9	very large LDL	25−30	1.019
L10	IDL	30−40	1.015
L11	small VLDL	40−60	1.010
L12	large VLDL	60−80	1.006
L13	remnants	80−100	1.000
L14	small chylomicrons	100−150	0.980
L15	large chylomicrons	> 150	0.960

Table A.2: Metabolites measured in KORA and TwinsUK. This Table summarises the super pathway, measurement platform, number of samples in KORA and TwinsUK for which we measured the metabolite (N KORA, N TwinsUK), the normalised minimal and maximal values and the median relative standard deviation (RSD) for each metabolite. The minimal and maximal value and the RSD are calculated from technical replicates of a pool of human plasma that has been well characterised at Metabolon (Evans *et al.*, 2009). "The biochemical identity of the metabolites is in general determined using adequate pure substances; in cases where metabolite identities were inferred based on their fragmentation spectrum and other biochemical evidence, these are indicated by a '*'" (Suhre *et al.*, 2011a).

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
2-aminobutyrate	amino acid	LC/MS pos	1775	1051	0.671	1.407	8.7
2-hydroxybutyrate (AHB)	amino acid	GC/MS	1775	1052	0.734	1.726	9.0
2-hydroxyisobutyrate	amino acid	GC/MS	1641	930	0.385	1.906	21.7
2-methylbutyroyl-carnitine	amino acid	LC/MS pos	1706	1027	0.348	2.217	17.2
3-(3-hydroxyphenyl)-propionate	amino acid	LC/MS neg	276	100			
3-(4-hydroxyphenyl)-lactate	amino acid	LC/MS neg	1770	1052	0.657	1.307	8.3
3-hydroxy-2-ethylpropionate	amino acid	GC/MS	894	0	0.484	1.465	18.9
3-indoxyl sulfate	amino acid	LC/MS neg	1774	1051	0.686	1.185	5.6
3-methoxytyrosine	amino acid	LC/MS pos	1468	379	0.598	2.452	21.3
3-methyl-2-oxobutyrate	amino acid	LC/MS neg	1776	1044	0.540	1.415	10.4
3-methyl-2-oxovalerate	amino acid	LC/MS neg	1776	1052	0.601	1.351	8.3
3-methylhistidine	amino acid	LC/MS neg	661	742	0.664	1.327	8.0
3-phenylpropionate (hydrocinnamate)	amino acid	LC/MS neg	1268	855	0.511	1.635	17.2
4-acetamidobutanoate	amino acid	LC/MS pos	1621	715	0.481	1.728	15.3
4-hydroxyphenylacetate	amino acid	GC/MS	388	0	0.462	1.438	17.9
4-methyl-2-oxopentanoate	amino acid	LC/MS neg	1776	1052	0.653	1.376	9.3
5-oxoproline	amino acid	LC/MS pos	1776	1052	0.713	1.190	5.7
alanine	amino acid	GC/MS	1775	1052	0.367	1.828	15.1

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
alpha-hydroxyisovalerate	amino acid	LC/MS neg	1776	1052	0.634	1.348	9.6
arginine	amino acid	LC/MS neg	1746	1017	0.309	1.648	14.5
asparagine	amino acid	GC/MS	1768	1045	0.310	1.918	25.5
aspartate	amino acid	GC/MS	1732	1049	0.228	2.220	25.9
beta-hydroxyisovalerate	amino acid	LC/MS neg	1595	992	0.437	1.995	18.9
betaine	amino acid	LC/MS pos	1775	1052	0.381	1.566	9.0
C-glycosyltryptophan*	amino acid	LC/MS pos	1774	1049	0.647	1.297	9.4
citrulline	amino acid	LC/MS pos	1767	1047	0.702	1.606	10.3
creatine	amino acid	LC/MS pos	1776	1052	0.796	1.122	5.4
creatinine	amino acid	LC/MS pos	1775	1052	0.683	1.556	10.8
cysteine	amino acid	GC/MS	1771	1013	0.197	1.769	22.9
cysteine-glutathione disulfide	amino acid	LC/MS pos	1638	367			
cystine	amino acid	GC/MS	1412	0			
dimethylarginine (SDMA + ADMA)	amino acid	LC/MS pos	1776	1052	0.599	2.190	14.0
glutamate	amino acid	GC/MS	1775	1052	0.459	2.058	15.7
glutamine	amino acid	LC/MS pos	1776	1052	0.772	1.248	7.0
glutaroyl carnitine	amino acid	LC/MS pos	1729	1012	0.588	1.600	11.8
glycine	amino acid	GC/MS	1775	1052	0.277	1.715	15.3
histidine	amino acid	LC/MS neg	1776	1052	0.790	1.218	5.4
homocitrulline	amino acid	LC/MS pos	1412	650	0.448	2.244	26.1
homostachydrine*	amino acid	LC/MS pos	1471	162	0.379	1.878	17.0
hydroxyisovaleroyl carnitine	amino acid	LC/MS pos	1530	960	0.241	3.771	34.6
indoleacetate	amino acid	LC/MS pos	1750	935	0.511	1.556	14.4
indolelactate	amino acid	LC/MS pos	1500	921	0.236	2.684	25.9
indolepropionate	amino acid	LC/MS pos	1775	1051	0.493	1.465	12.7
isobutyrylcarnitine	amino acid	LC/MS pos	1776	1049	0.575	1.391	8.8
isoleucine	amino acid	LC/MS pos	1776	1052	0.774	1.179	6.0
isovalerylcarnitine	amino acid	LC/MS pos	1762	1041	0.606	1.808	12.9
kynurenine	amino acid	LC/MS pos	1776	1052	0.706	1.303	7.8
leucine	amino acid	LC/MS pos	1776	1052	0.772	1.133	6.0

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
levulinate (4-oxovalerate)	amino acid	LC/MS pos or neg	1066	993	0.321	44.868	204.1
lysine	amino acid	LC/MS pos	1776	1052	0.457	1.700	13.5
methionine	amino acid	LC/MS neg	1776	1052	0.745	1.202	6.5
N-(2-furoyl)glycine	amino acid	LC/MS pos	429	102	0.385	1.925	22.3
N-acetylalanine	amino acid	LC/MS neg	1711	1051	0.642	1.764	11.7
N-acetylglycine	amino acid	GC/MS	1515	835	0.216	2.948	21.9
N-acetylornithine	amino acid	LC/MS pos	1762	1044	0.387	2.303	26.2
N-acetylthreonine	amino acid	LC/MS neg	1416	880	0.414	2.092	22.2
ornithine	amino acid	LC/MS pos	1776	1040	0.315	3.312	26.6
p-cresol sulfate	amino acid	LC/MS neg	1776	1052	0.756	1.152	3.6
phenol sulfate	amino acid	LC/MS neg	1776	1052	0.571	1.182	4.9
phenylacetate	amino acid	LC/MS neg	777	793	0.503	1.965	22.4
phenylacetylglutamine	amino acid	LC/MS pos	1776	1051	0.756	1.257	6.9
phenylalanine	amino acid	LC/MS pos	1776	1052	0.787	1.144	5.8
phenyllactate (PLA)	amino acid	LC/MS neg	1081	630	0.488	1.518	15.8
pipecolate	amino acid	LC/MS pos	1776	1052	0.687	1.303	7.4
proline	amino acid	LC/MS pos	1776	1052	0.791	1.211	5.9
pyroglutamine*	amino acid	LC/MS pos	1772	1051	0.721	1.455	9.3
serine	amino acid	GC/MS	1775	1052	0.372	2.164	17.1
serotonin (5HT)	amino acid	LC/MS pos	1758	1008			
stachydrine	amino acid	LC/MS pos	1775	1049	0.758	1.205	6.4
threonine	amino acid	LC/MS pos	1694	1039	0.623	1.371	10.1
tiglyl carnitine	amino acid	LC/MS pos	1339	0	0.405	2.357	24.3
trans-4-hydroxyproline	amino acid	GC/MS	1775	1051	0.615	2.152	12.5
tryptophan	amino acid	LC/MS pos	1776	1052	0.787	1.157	6.2
tyrosine	amino acid	LC/MS pos	1776	1052	0.745	1.156	6.1
urea	amino acid	GC/MS	1775	1052	0.510	1.481	10.1
valine	amino acid	LC/MS pos	1776	1052	0.792	1.139	5.9
1,5-anhydroglucitol (1,5-AG)	carbohydrate	LC/MS neg	1772	1051	0.755	1.288	5.6
1,6-anhydroglucose	carbohydrate	GC/MS	418	296	0.335	1.845	20.3
arabinose	carbohydrate	GC/MS	1011	567	0.358	1.800	21.0

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
arabitol	carbohydrate	GC/MS	1770	0	0.161	1.602	18.9
erythronate*	carbohydrate	GC/MS	1755	1032	0.519	2.225	15.9
erythrose	carbohydrate	GC/MS	1465	838	0.117	2.876	36.0
fructose	carbohydrate	GC/MS	1774	1052	0.021	2.763	26.7
glucose	carbohydrate	GC/MS	1775	1052	0.720	1.649	9.6
glycerate	carbohydrate	GC/MS	1762	1035	0.157	1.607	13.2
lactate	carbohydrate	GC/MS	1775	1052	0.734	1.885	8.6
mannitol	carbohydrate	GC/MS	1529	799	0.311	2.145	31.4
mannose	carbohydrate	GC/MS	1775	1052	0.109	1.995	17.1
pyruvate	carbohydrate	GC/MS	1735	984	0.196	3.223	37.1
threitol	carbohydrate	GC/MS	1640	898	0.358	2.077	23.5
alpha-tocopherol	cofactors and vitamins	GC/MS	1770	1042	0.324	2.796	19.3
ascorbate (Vitamin C)	cofactors and vitamins	GC/MS	1581	518	1.000	1.000	0.0
bilirubin (E,E)*	cofactors and vitamins	LC/MS pos	1776	1042	0.458	2.060	21.5
bilirubin (E,Z or Z,E)*	cofactors and vitamins	LC/MS pos	1489	775	0.301	3.148	35.0
bilirubin (Z,Z)	cofactors and vitamins	LC/MS neg	1728	976	0.666	2.738	31.0
biliverdin	cofactors and vitamins	LC/MS neg	1181	518	0.198	3.226	28.8
gamma-tocopherol	cofactors and vitamins	GC/MS	983	620	0.311	2.444	22.7
heme*	cofactors and vitamins	LC/MS pos	1702	818	0.431	2.066	24.3
pantothenate	cofactors and vitamins	LC/MS pos	1664	981	0.616	1.989	18.4
pyridoxate	cofactors and vitamins	LC/MS neg	1732	1045	0.539	1.592	13.4
riboflavin (Vitamin B2)	cofactors and vitamins	LC/MS pos	308	0			
threonate	cofactors and vitamins	GC/MS	1775	1047	0.215	3.799	18.7
trigonelline (N'-methyl-nicotinate)	cofactors and vitamins	LC/MS pos	745	0	0.637	1.634	13.9

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
acetylphosphate	energy	GC/MS	1775	1052	0.539	1.884	18.0
alpha-ketoglutarate	energy	GC/MS	1125	496	0.240	3.414	30.2
citrate	energy	GC/MS	1774	1052	0.614	1.757	9.2
malate	energy	GC/MS	1588	854	0.396	3.420	25.1
phosphate	energy	GC/MS	1775	1052	0.658	1.507	7.6
succinylcarnitine	energy	LC/MS pos	1545	786	0.613	1.837	15.9
1-arachidonoylglycero-phosphocholine*	lipid	LC/MS pos	1625	913	0.356	6.295	47.0
1-arachidonoylglycero-phosphoethanolamine*	lipid	LC/MS neg	1774	1051	0.253	5.307	38.2
1-arachidonoylglycero-phosphoinositol*	lipid	LC/MS neg	1770	1047	0.519	2.184	20.7
1-docosahexaenoyl-glycerophosphocholine*	lipid	LC/MS pos	1776	1040	0.244	6.081	55.4
1-eicosadienoylglycero-phosphocholine*	lipid	LC/MS pos	1648	711	0.317	6.934	62.3
1-eicosatrienoylglycero-phosphocholine*	lipid	LC/MS pos	1776	1051	0.083	5.910	42.1
1-heptadecanoyl-glycerophosphocholine	lipid	LC/MS pos	1741	882	0.302	10.327	71.2
1-linoleoylglycerol (1-monolinolein)	lipid	LC/MS neg	1766	1044	0.109	13.084	75.0
1-linoleoylglycero-phosphocholine	lipid	LC/MS pos	1775	1051	0.266	5.341	39.5
1-linoleoylglycero-phosphoethanolamine*	lipid	LC/MS neg	1776	1052	0.239	3.967	32.4
1-myristoylglycero-phosphocholine	lipid	LC/MS pos	1774	1051	0.244	4.290	43.8
1-oleoylglycerol (1-mono-olein)	lipid	LC/MS pos	1699	717	0.282	6.173	64.5
1-oleoylglycerophospho-choline	lipid	LC/MS pos	1776	1052	0.376	3.101	33.9
1-oleoylglycerophospho-ethanolamine	lipid	LC/MS neg	1745	1020	0.133	3.986	38.1
1-palmitoleoylglycero-phosphocholine*	lipid	LC/MS pos	1776	1052	0.312	5.059	45.6
1-palmitoylglycerol (1-monopalmitin)	lipid	GC/MS	1617	927	0.269	2.370	19.9

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
1-palmitoylglycero-phosphocholine	lipid	LC/MS pos	1776	1052	0.310	2.413	24.4
1-palmitoylglycero-phosphoethanolamine	lipid	LC/MS neg	1760	1032	0.235	3.144	28.8
1-palmitoylglycero-phosphoinositol*	lipid	LC/MS neg	1512	907	0.564	2.162	18.9
1-stearoylglycerol (1-monostearin)	lipid	GC/MS	1418	794	0.244	1.972	23.5
1-stearoylglycero-phosphocholine	lipid	LC/MS pos	1776	1052	0.476	4.345	37.6
1-stearoylglycero-phosphoethanolamine	lipid	LC/MS neg	1578	875	0.177	3.313	34.8
1-stearoylglycero-phosphoinositol	lipid	LC/MS neg	1748	1036	0.345	2.712	22.8
2-hydroxypalmitate	lipid	LC/MS neg	1776	1052	0.384	1.676	13.8
2-hydroxystearate	lipid	LC/MS neg	1771	1015	0.317	1.637	16.2
2-linoleoylglycero-phosphocholine*	lipid	LC/MS pos	1688	1015	0.333	4.841	41.2
2-linoleoylglycero-phosphoethanolamine*	lipid	LC/MS neg	1078	0	0.221	3.793	42.4
2-oleoylglycerophospho-choline*	lipid	LC/MS pos	1770	1050	0.266	4.162	34.8
2-palmitoylglycero-phosphocholine*	lipid	LC/MS pos	1776	1052	0.118	4.967	40.5
2-stearoylglycero-phosphocholine*	lipid	LC/MS pos	1775	1046	0.417	11.443	68.7
2-tetradecenoyl carnitine	lipid	LC/MS pos	1649	741	0.445	2.608	30.4
3-carboxy-4-methyl-5-propyl-2-furanpropano-ate (CMPF)	lipid	LC/MS neg	1768	1051	0.626	1.338	9.2
3-dehydrocarnitine*	lipid	LC/MS pos	1776	1052	0.693	2.317	10.9
3-hydroxybutyrate (BHBA)	lipid	GC/MS	1775	1052	0.668	1.444	8.0
5-dodecenoate (12 : 1n7)	lipid	LC/MS neg	1762	1047	0.496	1.766	14.5
7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca)	lipid	LC/MS neg	1776	1050	0.497	4.109	29.2
10-heptadecenoate (17 : 1n7)	lipid	LC/MS neg	1776	1052	0.650	1.868	10.2

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
10-nonadecenoate (19 : 1n9)	lipid	LC/MS neg	1776	1052	0.464	1.694	12.3
10-undecenoate (11 : 1n1)	lipid	LC/MS neg	1776	1049	0.516	1.680	17.1
acetylcarnitine	lipid	LC/MS pos	1776	1052	0.575	1.317	8.5
adrenate (22 : 4n6)	lipid	LC/MS neg	1776	1052	0.530	2.854	27.0
androsterone sulfate	lipid	LC/MS neg	1772	1049	0.679	1.215	6.9
arachidonate (20 : 4n6)	lipid	LC/MS neg	1776	1052	0.688	1.245	6.7
butyrylcarnitine	lipid	LC/MS pos	1774	1051	0.635	1.977	14.0
caprate (10 : 0)	lipid	LC/MS neg	1776	1051	0.789	1.271	6.9
caproate (6 : 0)	lipid	LC/MS neg	1776	1052	0.636	1.839	13.3
caprylate (8 : 0)	lipid	LC/MS neg	1776	1052	0.703	1.423	9.8
carnitine	lipid	LC/MS pos	1776	1052	0.804	1.203	5.7
cholate	lipid	LC/MS neg	1214	790	0.285	1.786	22.0
cholesterol	lipid	GC/MS	1775	1052	0.720	1.749	9.7
cortisol	lipid	LC/MS pos	1773	1051	0.770	1.311	7.5
cortisone	lipid	LC/MS pos	1730	910	0.480	1.599	17.4
choline	lipid	LC/MS pos	1775	1052	0.747	1.215	7.5
decanoylcarnitine	lipid	LC/MS pos	1776	1052	0.585	1.473	12.3
dehydroisoandrosterone sulfate (DHEA-S)	lipid	LC/MS neg	1776	1052	0.583	1.142	4.2
deoxycholate	lipid	LC/MS neg	1455	750	0.378	2.286	25.7
dihomo-linoleate (20 : 2n6)	lipid	LC/MS neg	1776	1052	0.609	1.673	11.1
dihomo-linolenate (20 : 3n3 or n6)	lipid	LC/MS neg	1776	1052	0.580	1.588	9.4
docosahexaenoate (DHA; 22 : 6n3)	lipid	LC/MS neg	1776	1052	0.662	1.288	8.5
docosapentaenoate (n3 DPA; 22 : 5n3)	lipid	LC/MS neg	1776	1052	0.527	1.694	11.8
dodecanedioate	lipid	LC/MS neg	946	898	0.451	1.853	20.7
eicosapentaenoate (EPA; 20 : 5n3)	lipid	LC/MS neg	1776	1052	0.568	1.475	9.7
eicosenoate (20 : 1n9 or 11)	lipid	LC/MS neg	1776	1052	0.571	1.487	12.9

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
epiandrosterone sulfate	lipid	LC/MS neg	1773	1049	0.645	1.218	5.4
estrone 3-sulfate	lipid	LC/MS neg	261	105	0.784	1.580	12.8
glycerol	lipid	GC/MS	1775	1052	0.011	1.447	8.9
glycerol 3-phosphate (G3P)	lipid	GC/MS	1775	1050	0.374	2.561	15.6
glycerophosphorylcholine (GPC)	lipid	LC/MS pos	1772	1049	0.239	2.722	14.6
glycochenodeoxycholate	lipid	LC/MS neg	1576	984	0.531	1.569	14.1
glycocholate	lipid	LC/MS pos	1168	685	0.678	1.395	10.1
glycodeoxycholate	lipid	LC/MS neg	874	609	0.429	1.762	18.8
heptanoate (7 : 0)	lipid	LC/MS neg	1775	1052	0.639	1.671	12.2
hexadecanedioate	lipid	LC/MS neg	1022	906	0.530	2.104	20.0
hexanoylcarnitine	lipid	LC/MS pos	1776	1044	0.572	1.502	12.6
hyodeoxycholate	lipid	LC/MS neg	1314	640	0.525	1.428	13.6
inositol 1-phosphate (I1P)	lipid	GC/MS	1391	0			
isovalerate	lipid	LC/MS neg	1730	1014	0.488	1.481	12.8
lathosterol	lipid	GC/MS	841	425	0.381	1.713	20.8
laurate (12 : 0)	lipid	LC/MS neg	1776	1052	0.783	1.626	10.6
laurylcarnitine	lipid	LC/MS pos	1578	765	0.286	4.025	29.2
linoleamide (18 : 2n6)	lipid	LC/MS pos	1776	0			
linoleate (18 : 2n6)	lipid	LC/MS neg	1776	1052	0.718	1.255	6.1
linolenate [alpha or gamma; (18 : 3n3 or 6)]	lipid	LC/MS neg	1776	1052	0.693	1.316	8.2
margarate (17 : 0)	lipid	LC/MS neg	1776	1052	0.626	1.565	11.3
myo-inositol	lipid	GC/MS	1775	1052	0.391	1.625	14.3
myristate (14 : 0)	lipid	LC/MS neg	1776	1052	0.741	1.296	7.1
myristoleate (14 : 1n5)	lipid	LC/MS neg	1776	1052	0.741	1.281	7.8
n-Butyl Oleate	lipid	GC/MS	1374	695	0.479	1.746	18.9
nonadecanoate (19 : 0)	lipid	LC/MS neg	1767	1041	0.448	1.841	18.6
octadecanedioate	lipid	LC/MS neg	1513	941	0.258	2.321	27.5
octanoylcarnitine	lipid	LC/MS pos	1776	1052	0.643	1.390	10.0
oleamide	lipid	LC/MS pos	1776	0	0.234	11.500	83.4

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
oleate (18 : 1n9)	lipid	LC/MS neg	1776	1052	0.735	1.257	6.7
oleoylcarnitine	lipid	LC/MS pos	1772	1042	0.289	2.189	24.3
palmitate (16 : 0)	lipid	LC/MS neg	1776	1052	0.717	1.363	7.8
palmitoleate (16 : 1n7)	lipid	LC/MS neg	1776	1052	0.737	1.454	7.5
palmitoylcarnitine	lipid	LC/MS pos	1763	1032	0.297	2.473	28.6
pelargonate (9 : 0)	lipid	LC/MS neg	1776	1052	0.701	1.337	8.5
pentadecanoate (15 : 0)	lipid	GC/MS	1716	1018	0.397	2.978	21.9
propionylcarnitine	lipid	LC/MS pos	1776	1052	0.688	1.488	10.8
scyllo-inositol	lipid	GC/MS	1511	897	0.475	1.831	28.1
sebacate (decanedioate)	lipid	LC/MS neg	400	0			
stearate (18 : 0)	lipid	LC/MS neg	1776	1052	0.691	1.302	8.8
stearidonate (18 : 4n3)	lipid	LC/MS neg	1769	1050	0.546	1.463	13.3
stearoylcarnitine	lipid	LC/MS pos	1607	841	0.349	3.503	29.4
taurochenodeoxycholate	lipid	LC/MS neg	1051	611	0.191	1.834	20.5
taurocholate	lipid	LC/MS neg	706	488	0.405	2.616	26.9
taurodeoxycholate	lipid	LC/MS neg	968	601	0.372	1.814	21.2
taurolithocholate 3-sulfate	lipid	LC/MS neg	1592	959	0.178	2.873	27.4
tetradecanedioate	lipid	LC/MS neg	662	532			
thromboxane B2	lipid	LC/MS neg	1752	1031	0.428	1.386	7.6
undecanoate (11 : 0)	lipid	LC/MS neg	1757	1037	0.541	2.361	16.9
ursodeoxycholate	lipid	LC/MS neg	962	674	0.429	1.814	16.6
valerate	lipid	LC/MS neg	1440	742	0.406	2.463	28.0
7-methylguanine	nucleotide	LC/MS pos	1665	978	0.434	3.013	33.4
adenosine	nucleotide	LC/MS pos	411	0			
allantoin	nucleotide	GC/MS	742	583	0.455	2.672	27.9
guanosine	nucleotide	LC/MS pos	1655	697			
hypoxanthine	nucleotide	LC/MS neg	1762	1030	0.280	1.660	16.8
inosine	nucleotide	LC/MS neg	1739	944			
N1-methyladenosine	nucleotide	LC/MS pos	1775	1052	0.672	1.543	12.0
N2,N2-dimethylguanosine	nucleotide	LC/MS pos	825	261	0.586	4.034	28.9
pseudouridine	nucleotide	LC/MS pos	1776	1052	0.600	1.797	14.6

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
urate	nucleotide	LC/MS neg	1776	1052	0.680	1.175	4.9
uridine	nucleotide	LC/MS neg	1776	1052	0.743	1.205	7.4
xanthine	nucleotide	LC/MS pos	1771	1052	0.253	2.193	22.1
ADpSGEGDFXAEGG-GVR*	peptide	LC/MS pos	1773	1045	0.160	1.708	16.4
ADSGEGDFXAEGGGVR*	peptide	LC/MS pos	1776	1052	0.597	1.319	10.9
aspartylphenylalanine	peptide	LC/MS pos	1758	1050	0.746	1.090	12.9
bradykinin, des-arg(9)	peptide	LC/MS pos	1504	819	0.438	1.303	10.3
DSGEGDFXAEGGGVR*	peptide	LC/MS pos	1775	1051	0.372	1.838	20.4
gamma-glutamyl-glutamate	peptide	LC/MS pos	654	285	0.478	2.202	20.1
gamma-glutamyl-glutamine	peptide	LC/MS pos	1776	1052	0.533	1.557	15.3
gamma-glutamyl-isoleucine*	peptide	LC/MS pos	1023	373	0.558	2.185	16.8
gamma-glutamylleucine	peptide	LC/MS pos	1776	1051	0.650	1.484	10.9
gamma-glutamyl-methionine*	peptide	LC/MS pos	1384	862	0.414	2.321	23.1
gamma-glutamyl-phenylalanine	peptide	LC/MS pos	1761	1010	0.425	1.674	17.0
gamma-glutamyl-threonine*	peptide	LC/MS pos	1345	476	0.268	1.900	21.5
gamma-glutamyltyrosine	peptide	LC/MS pos	1677	903	0.567	2.110	17.2
gamma-glutamylvaline	peptide	LC/MS pos	1755	1033	0.726	1.429	9.3
glycylvaline	peptide	LC/MS pos	1215	905			
HWESASXX*	peptide	LC/MS pos	1722	1052	0.552	1.365	11.4
leucylleucine	peptide	LC/MS pos	1060	968	0.716	1.312	8.7
pro-hydroxy-pro	peptide	LC/MS pos	1774	1052	0.400	2.035	20.1
pyroglutamylglycine	peptide	LC/MS neg	793	798	0.790	1.210	8.2
1,3,7-trimethylurate	xenobiotics	LC/MS neg	314	483			
1,7-dimethylurate	xenobiotics	LC/MS neg	1025	805	0.651	1.289	8.4
1-methylurate	xenobiotics	LC/MS pos	1113	553	0.584	2.236	22.9
1-methylxanthine	xenobiotics	LC/MS pos	1184	606	0.923	1.077	7.7
2-methoxyacetamino-phen sulfate*	xenobiotics	LC/MS neg	26	187			

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
2-hydroxyacetaminophen sulfate*	xenobiotics	LC/MS neg	76	328	0.813	1.314	9.2
2-hydroxyhippurate (salicylurate)	xenobiotics	LC/MS neg	355	351	0.738	1.262	19.2
3-(cystein-S-yl)acetaminophen*	xenobiotics	LC/MS pos	18	165			
3-ethylphenylsulfate*	xenobiotics	LC/MS neg	166	0	0.599	2.585	24.9
3-methylxanthine	xenobiotics	LC/MS pos	1127	533	0.415	2.168	24.8
4-acetamidophenol	xenobiotics	GC/MS	0	143	1.000	1.000	0.0
4-acetaminophen sulfate	xenobiotics	LC/MS neg	122	376	0.667	1.566	8.6
4-ethylphenylsulfate	xenobiotics	LC/MS neg	1423	866	0.541	1.820	14.2
4-vinylphenol sulfate	xenobiotics	LC/MS neg	1733	1001	0.684	1.591	6.5
7-methylxanthine	xenobiotics	LC/MS pos	1319	581	0.422	3.144	33.2
benzoate	xenobiotics	LC/MS neg	1763	1052	0.673	1.378	9.7
caffeine	xenobiotics	LC/MS pos	1721	1038	0.591	1.458	12.2
catechol sulfate	xenobiotics	LC/MS neg	1776	1052	0.745	1.232	4.9
cotinine	xenobiotics	LC/MS pos	284	126	0.607	1.611	15.7
erythritol	xenobiotics	GC/MS	1772	1048	0.371	1.604	15.1
glycerol 2-phosphate	xenobiotics	GC/MS	1237	559	0.255	3.184	21.5
hippurate	xenobiotics	LC/MS pos	1766	1051	0.603	1.607	15.0
hydroquinone sulfate	xenobiotics	LC/MS neg	354	103	0.928	1.072	7.2
hydroxypioglitazone*	xenobiotics	LC/MS pos	6	2			
ibuprofen	xenobiotics	LC/MS neg	25	66	0.618	2.037	17.3
metoprolol	xenobiotics	LC/MS pos	69	1			
metoprolol acid metabolite*	xenobiotics	LC/MS pos	149	57			
naproxen	xenobiotics	LC/MS neg	2	7			
p-acetamidophenyl-glucuronide	xenobiotics	LC/MS pos	60	231	0.524	2.727	25.9
paraxanthine	xenobiotics	LC/MS pos	1667	969	0.544	1.645	13.2
pioglitazone*	xenobiotics	LC/MS pos	6	2			
piperine	xenobiotics	LC/MS pos	1746	966	0.514	1.765	15.7
quinate	xenobiotics	GC/MS	1460	737	0.265	1.978	23.9
saccharin	xenobiotics	LC/MS neg	410	322			

Table A.2 (cont.)

metabolite	super pathway	measurement platform	N KORA	N TwinsUK	Min Value	Max Value	RSD (%)
salicylate	xenobiotics	GC/MS	484	197	0.475	1.616	19.8
salicyluric glucuronide*	xenobiotics	LC/MS neg	272	99	1.000	1.000	0.0
theobromine	xenobiotics	LC/MS pos	1755	1042	0.730	1.331	8.3
theophylline	xenobiotics	LC/MS neg	1653	977	0.625	1.725	15.8
thymol sulfate	xenobiotics	LC/MS neg	1064	626	0.590	1.554	10.0
carbamazepine*		LC/MS pos	5	4			

Table A.3: One hundred and one SNPs published by Teslovich *et al.* (2010). This Table summarises the 101 SNPs that were extracted from the KORA genotype data for the application of the candidate locus approach together with proxy SNPs for the replication of significant associations in the GRAPHIC study.

gene	SNP	chr	position	proxy	position	R^2	D'
LDLRAP1	rs12027135	1	25648320				
PABPC4	rs4660293	1	39800767				
PCSK9	rs2479409	1	55277238				
ANGPTL3	rs2131925	1	62823186				
EVI5	rs7515577	1	92782026				
SORT1	rs629301	1	109619829	rs646776	109620053	1	1
ZNF648	rs1689800	1	180435508				
MOSC1	rs2642442	1	219037216				
GALNT2	rs4846914	1	228362314				
IRF2BP	rs514230	1	232925220				
APOB	rs1367117	2	21117405				
APOB	rs1042034	2	21078786				
GCKR	rs1260326	2	27584444				
ABCG5/8	rs4299376	2	43926080				
RAB3GAP1	rs7570971	2	136039146				
COBLL1	rs12328675	2	165249046				
COBLL1	rs10195252	2	165221337				
IRS1	rs2972146	2	226837161				
RAF1	rs2290159	3	12603920				
MSL2L1	rs645040	3	137409312				
KLHL8	rs442177	4	88249285				
SLC39A8	rs13107325	4	103407732				
ARL15	rs6450176	5	53333782				
MAP3K1	rs9686661	5	55897543				
HMGCR	rs12916	5	74692295				
TIMD4	rs6882076	5	156322875				
MYLIP	rs3757354	6	16235386				
HFE	rs1800562	6	26201120				
HLA	rs3177928	6	32520413				

Table A.3 (cont.)

gene	SNP	chr	position	proxy	position	R^2	D'
HLA	rs2247056	6	31373469				
C6orf106	rs2814982	6	34654538				
C6orf106	rs2814944	6	34660775				
FRK	rs9488822	6	116419586				
CITED2	rs605066	6	139871359				
LPA	rs1564348	6	160498850				
LPA	rs1084651	6	161009807				
DNAH11	rs12670798	7	21549442				
NPC1L1	rs2072183	7	44545705				
TYW1B	rs13238203	7	71767603				
MLXIPL	rs17145738	7	72620810				
KLF14	rs4731702	7	130083924				
PPP1R3B	rs9987289	8	9222556				
PINX1	rs11776767	8	10721339				
NAT2	rs1495741	8	18299989				
LPL	rs12678919	8	19888502				
CYP7A1	rs2081687	8	59474251				
TRPS1	rs2737229	8	116717740				
TRPS1	rs2293889	8	116668374				
TRIB1	rs2954029	8	126551803				
PLEC1	rs11136341	8	145115531				
TTC39B	rs581080	9	15295378				
ABCA1	rs1883025	9	106704122				
JMJD1C	rs10761731	10	64697616				
CYP26A1	rs2068888	10	94829632				
GPAM	rs2255141	10	113923876				
AMPD3	rs2923084	11	10345358				
SPTY2D1	rs10128711	11	18620817				
LRP4	rs3136441	11	46699823				
FADS1-2-3	rs174546	11	61328054	rs102275	61314379	1	1
APOA1	rs964184	11	116154127	rs12286037	116157417	0.588	1
UBASH3B	rs7941030	11	122027585				
ST3GAL4	rs11220462	11	125753421				
PDE3A	rs7134375	12	20365025				

Table A.3 (cont.)

gene	SNP	chr	position	proxy	position	R^2	D'
LRP1	rs11613352	12	56130316				
MVK	rs7134594	12	108484576				
BRAP	rs11065987	12	110556807				
HNF1A	rs1169288	12	119901033				
SBNO1	rs4759375	12	122362191				
ZNF664	rs4765127	12	123026120				
SCARB1	rs838880	12	123827546				
NYNRIN	rs8017377	14	23952898				
CAPN3	rs2412710	15	40471079				
FRMD5	rs2929282	15	42033223				
LIPC	rs1532085	15	56470658	rs4775041	56461987	0.536	0.904
LACTB	rs2652834	15	61183920				
CTF1	rs11649653	16	30825988				
CETP	rs3764261	16	55550825				
LCAT	rs16942887	16	66485543				
HPR	rs2000999	16	70665594				
CMIP	rs2925979	16	80092291				
STARD3	rs11869286	17	35063744				
OSBPL7	rs7206971	17	42780114				
ABCA8	rs4148008	17	64386889				
PGS1	rs4129767	17	73889077				
LIPG	rs7241918	18	45418715				
MC4R	rs12967135	18	56000003				
ANGPTL4	rs7255436	19	8339196				
LDLR	rs6511720	19	11063306				
LOC55908	rs737337	19	11208493				
CILP2	rs10401969	19	19268718				
APOE	rs4420638	19	50114786				
APOE	rs439401	19	50106291				
FLJ36070	rs492602	19	53898229				
LILRA3	rs386000	19	59484573				
ERGIC3	rs2277862	20	33616196				
MAFB	rs2902940	20	38524901				
TOP1	rs6029526	20	39244689				

Table A.3 (cont.)

gene	SNP	chr	position	proxy	position	R^2	D'
HNF4A	rs1800961	20	42475778				
PLTP	rs6065906	20	43987422	rs6073952	43970339	0.877	1
UBE2L3	rs181362	22	20262068				
PLA2G6	rs5756931	22	36875979				

Table A.4: Summary statistics of lipoprotein subfractions. Means and standard deviations of age, BMI, serum lipids (HDL-C, LDL-C, TC, TG) and lipoprotein subfractions (L1-L15) were summarised for each cohort. Statistics were calculated for males and females separately. In addition, a stratification by parents and offspring was done in GRAPHIC. [a] The units for the serum lipids are mg/dl for KORA and mmol/l for GRAPHIC. The conversion factor is [mmol/l] * 38.67 = [mg/dl] for cholesterol and [mmol/l] * 87.5 = [mg/dl] for TG (Petersen *et al.*, 2012).

	KORA		GRAPHIC				HuMet fasting	HuMet lipid tolerance test
	males (N=873)	females (N=918)	fathers (N=497)	mothers (N=488)	sons (N=491)	daughters (N=472)	males (N=15)	males (N=15)
age (years)	61.1 (8.8)	60.5 (8.8)	53.8 (4.3)	51.8 (4.4)	25.1 (5.1)	25.96 (5.4)	27.8 (2.08)	
BMI (kg/m^2)	28.4 (4.3)	27.9 (5.3)	27.8 (4.0)	27.0 (4.5)	24.9 (4.1)	24.6 (4.9)	23.1 (1.76)	
HDL-C[a]	51.0 (12.7)	61.8 (14.4)	1.32 (0.30)	1.64 (0.39)	1.3 (0.28)	1.47 (0.36)	NA	NA
LDL-C[a]	138.0 (34.2)	141.9 (35.6)	NA	NA	NA	NA	NA	NA
TC[a]	215.7 (38.7)	227.5 (38.9)	5.58 (0.99)	5.68 (0.97)	4.52 (0.89)	4.5 (0.82)	NA	NA
TG[a]	149.7 (111.5)	115.8 (69.0)	NA	NA	NA	NA	NA	NA
L1 (nmol/l)	23450.5 (3848.5)	22480.4 (3890.7)	20390.6 (3107.6)	18988.8 (3324)	19116.4 (3083.9)	18181.1 (3124.4)	18054.5 (3624.9)	18760.2 (3070.3)
L2 (nmol/l)	4003.4 (1773.3)	5356.7 (1702.2)	3525.3 (1349.4)	4983.9 (1483.3)	3466.4 (1415.2)	4447.2 (1611.3)	3515.7 (1006.3)	3136.2 (852.6)
L3 (nmol/l)	1468.9 (729.2)	2173.8 (916.5)	1407.6 (552.8)	2068.8 (735.7)	1367.5 (528.6)	1791.1 (692.6)	1439.7 (588.7)	1356.0 (521.7)
L4 (nmol/l)	1075.9 (175.4)	1184.9 (181.4)	883.02 (144.3)	968.6 (146.2)	750.8 (108.6)	838.5 (151.04)	746.8 (124.4)	755.8 (132.4)
L5 (nmol/l)	338.7 (111.2)	436.8 (126.5)	283.9 (79.2)	366.8 (93.8)	238.3 (67.8)	298.3 (90.2)	238.3 (84.4)	227.2 (73.3)
L6 (nmol/l)	329.4 (65.4)	352.2 (61.0)	275.4 (52.7)	299 (52.9)	222.4 (40)	246.4 (49.8)	220.0 (40.4)	210.5 (42.2)
L7 (nmol/l)	254.9 (69.9)	303.3 (66.7)	218.02 (50.2)	252.04 (52.5)	175.6 (37.4)	202.3 (46.7)	164.1 (45.7)	178.8 (47.5)
L8 (nmol/l)	228.6 (102.5)	207.7 (77.8)	196.96 (62.7)	182.3 (52.5)	147.5 (48.8)	142.15 (40.6)	140.1 (36.4)	118.1 (40.6)
L9 (nmol/l)	189.8 (88.0)	175.9 (64.4)	149.6 (54.1)	135.02 (43.7)	109.8 (43.8)	102.01 (31.8)	92.9 (29.9)	109.6 (32.9)
L10 (nmol/l)	114.4 (57.4)	96.2 (42.5)	104.3 (40)	86.7 (33.5)	78.7 (32.6)	66.04 (25.6)	67.4 (22.4)	71.8 (26.9)
L11 (nmol/l)	76.7 (50.67)	58.5 (36.3)	60.7 (32.2)	44.9 (23.9)	43.4 (24.9)	32.3 (16.5)	35.8 (17.3)	40.1 (17.5)
L12 (nmol/l)	14.0 (10.9)	10.3 (6.5)	13.3 (7.5)	9.7 (5.1)	9.97 (5.9)	7.3 (3.7)	7.2 (3.1)	11.2 (6.3)
L13 (nmol/l)	1.2 (0.98)	0.90 (0.73)	1.0 (0.69)	0.65 (0.48)	0.69 (0.5)	0.45 (0.32)	0.56 (0.35)	0.46 (0.29)
L14 (nmol/l)	0.59 (0.60)	0.39 (0.35)	0.57 (0.42)	0.38 (0.26)	0.42 (0.32)	0.27 (0.18)	0.27 (0.16)	0.48 (0.32)
L15 (nmol/l)	0.047 (0.06)	0.037 (0.02)	0.068 (0.06)	0.04 (0.03)	0.049 (0.05)	0.03 (0.02)	0.026 (0.01)	0.036 (0.03)

Table A.5: Quantiles of the p-gain density. Reported are the quantiles for various combinations of correlation values among the metabolic traits M_1, M_2 and M_3 with commonly $M_3 = M_1/M_2$. Note that some correlation settings are not possible if $M_3 = M_1/M_2$. Nevertheless, we extended the simulation analysis of the p-gain density to these correlation settings for the reason of completeness. In addition, we provided the quantiles for the simulated (sim) and calculated (calc) densities for the idealised case of fully correlated metabolic traits which are uncorrelated with the third metabolic trait.

correlation			quantiles								
(M_1;M_2)	(M_1;M_3)	(M_2;M_3)	1 %	2.5 %	5 %	10 %	50 %	90 %	95 %	97.5 %	99 %
0	0	0	34.13	13.18	6.59	3.30	0.63	0.10	0.05	0.02	0.01
0	0	±0.2	32.25	12.45	6.48	3.25	0.64	0.11	0.05	0.03	0.01
0	0	±0.4	25.57	10.76	5.80	3.04	0.64	0.11	0.06	0.03	0.01
0	0	±0.6	17.82	8.28	4.71	2.69	0.65	0.13	0.07	0.04	0.02
0	0	±0.8	8.54	4.97	3.25	2.10	0.67	0.16	0.09	0.05	0.02
0	0	±1	1.00	1.00	1.00	1.00	0.99	0.20	0.10	0.05	0.02
0	±0.2	±0.2	28.46	11.86	6.21	3.15	0.64	0.11	0.05	0.03	0.01
0	±0.2	±0.4	23.96	10.44	5.61	2.97	0.64	0.12	0.06	0.03	0.01
0	±0.2	±0.6	15.63	7.64	4.48	2.57	0.65	0.13	0.07	0.04	0.02
0	±0.2	±0.8	7.83	4.61	3.04	2.01	0.67	0.16	0.09	0.05	0.02
0	±0.4	±0.4	18.27	8.44	4.78	2.70	0.64	0.12	0.07	0.03	0.01
0	±0.4	±0.6	12.15	6.42	3.82	2.32	0.65	0.14	0.08	0.04	0.02
0	±0.4	±0.8	5.82	3.61	2.53	1.78	0.68	0.18	0.10	0.06	0.03
0	±0.6	±0.6	7.80	4.50	2.98	1.94	0.66	0.16	0.09	0.06	0.03
0	±0.6	±0.8	3.27	2.26	1.73	1.34	0.72	0.21	0.13	0.08	0.04
0	±0.8	±0.8	3.16	2.23	1.72	1.34	0.71	0.22	0.13	0.08	0.04
±0.2	0	0	32.68	13.20	6.53	3.30	0.64	0.10	0.05	0.02	0.01
±0.2	0	±0.2	31.44	12.59	6.50	3.27	0.64	0.11	0.05	0.03	0.01
±0.2	0	±0.4	25.23	11.19	5.81	3.07	0.64	0.12	0.06	0.03	0.01
±0.2	0	±0.6	16.49	8.05	4.62	2.65	0.65	0.13	0.07	0.04	0.02
±0.2	0	±0.8	8.64	4.92	3.24	2.11	0.68	0.16	0.09	0.05	0.02
±0.2	±0.2	±0.2	29.32	12.72	6.35	3.23	0.64	0.11	0.05	0.03	0.01
±0.2	±0.2	±0.4	24.63	10.72	5.61	2.96	0.64	0.12	0.06	0.03	0.01
±0.2	±0.2	±0.6	16.02	7.95	4.61	2.64	0.65	0.13	0.07	0.04	0.02
±0.2	±0.2	±0.8	8.25	4.82	3.15	2.08	0.67	0.16	0.09	0.05	0.02

Table A.5 (cont.)

correlation			quantiles								
$(M_1;M_2)$	$(M_1;M_3)$	$(M_2;M_3)$	1 %	2.5 %	5 %	10 %	50 %	90 %	95 %	97.5 %	99 %
±0.2	±0.2	±1	1.00	1.00	1.00	.001	0.99	0.21	0.11	0.06	0.02
±0.2	±0.4	±0.4	21.08	9.46	5.20	2.84	0.65	0.12	0.07	0.03	0.02
±0.2	±0.4	±0.6	14.16	7.00	4.17	2.45	0.65	0.14	0.08	0.04	0.02
±0.2	±0.4	±0.8	7.19	4.24	2.86	1.93	0.67	0.17	0.10	0.05	0.02
±0.2	±0.6	±0.6	9.65	5.22	3.34	2.12	0.66	0.16	0.09	0.05	0.03
±0.2	±0.6	±0.8	4.66	3.01	2.19	1.60	0.67	0.20	0.12	0.07	0.04
±0.2	±0.8	±0.8	2.48	1.85	1.49	1.22	0.71	0.25	0.16	0.11	0.06
±0.4	0	0	33.47	13.62	6.85	3.42	0.65	0.11	0.05	0.03	0.01
±0.4	0	±0.2	32.43	12.73	6.72	3.38	0.66	0.11	0.06	0.03	0.01
±0.4	0	±0.4	25.58	11.09	5.90	3.13	0.66	0.12	0.06	0.03	0.01
±0.4	0	±0.6	17.22	8.18	4.66	2.69	0.68	0.13	0.07	0.04	0.02
±0.4	0	±0.8	7.77	4.58	3.05	2.04	0.71	0.17	0.09	0.05	0.02
±0.4	±0.2	±0.2	29.70	12.35	6.41	3.30	0.66	0.11	0.06	0.03	0.01
±0.4	±0.2	±0.4	25.85	11.19	5.92	3.10	0.66	0.12	0.06	0.03	0.01
±0.4	±0.2	±0.6	18.47	8.57	4.87	2.75	0.67	0.14	0.07	0.04	0.02
±0.4	±0.2	±0.8	8.76	5.01	3.28	2.15	0.70	0.17	0.09	0.05	0.02
±0.4	±0.4	±0.4	21.56	9.93	5.38	2.91	0.66	0.13	0.07	0.04	0.01
±0.4	±0.4	±0.6	14.84	7.62	4.38	2.58	0.67	0.14	0.08	0.04	0.02
±0.4	±0.4	±0.8	8.46	4.84	3.16	2.07	0.69	0.18	0.10	0.06	0.03
±0.4	±0.4	±1	1.00	1.00	1.00	1.00	1.00	0.23	0.12	0.07	0.03
±0.4	±0.6	±0.6	11.44	6.04	3.72	2.28	0.66	0.16	0.09	0.05	0.03
±0.4	±0.6	±0.8	6.16	3.80	2.61	1.81	0.68	0.20	0.12	0.07	0.04
±0.4	±0.8	±0.8	3.12	2.24	1.75	1.37	0.69	0.25	0.17	0.11	0.07
±0.6	0	0	35.09	14.02	7.15	3.54	0.70	0.12	0.06	0.03	0.01
±0.6	0	±0.2	32.62	13.16	6.79	3.45	0.70	0.12	0.06	0.03	0.01
±0.6	0	±0.4	25.85	11.29	5.97	3.17	0.70	0.13	0.06	0.03	0.01
±0.6	0	±0.6	15.78	7.87	4.64	2.68	0.73	0.15	0.08	0.04	0.02
±0.6	0	±0.8	6.39	3.90	2.69	1.88	0.80	0.18	0.09	0.05	0.02
±0.6	±0.2	±0.2	31.86	13.11	6.83	3.46	0.69	0.12	0.06	0.03	0.01
±0.6	±0.2	±0.4	27.85	11.48	6.14	3.23	0.70	0.13	0.07	0.03	0.01
±0.6	±0.2	±0.6	17.86	8.71	5.01	2.86	0.71	0.15	0.08	0.04	0.02
±0.6	±0.2	±0.8	8.27	4.83	3.24	2.15	0.75	0.18	0.09	0.05	0.02
±0.6	±0.4	±0.4	25.18	10.85	5.74	3.13	0.70	0.13	0.07	0.04	0.02
±0.6	±0.4	±0.6	16.80	8.23	4.79	2.73	0.70	0.15	0.08	0.04	0.02

Table A.5 (cont.)

correlation			quantiles								
$(M_1;M_2)$	$(M_1;M_3)$	$(M_2;M_3)$	1 %	2.5 %	5 %	10 %	50 %	90 %	95 %	97.5 %	99 %
±0.6	±0.4	±0.8	8.76	5.14	3.36	2.20	0.73	0.18	0.10	0.06	0.02
±0.6	±0.6	±0.6	13.19	6.79	4.18	2.51	0.70	0.16	0.09	0.05	0.03
±0.6	±0.6	±0.8	7.54	4.51	3.00	2.01	0.71	0.20	0.12	0.07	0.04
±0.6	±0.6	±1	1.00	1.00	1.00	1.00	1.00	0.27	0.16	0.09	0.04
±0.6	±0.8	±0.8	4.50	3.01	2.23	1.63	0.71	0.25	0.17	0.11	0.07
±0.8	0	0	38.42	15.56	7.75	3.92	0.76	0.13	0.06	0.03	0.01
±0.8	0	±0.2	35.55	14.04	7.21	3.70	0.76	0.13	0.07	0.03	0.01
±0.8	0	±0.4	25.56	11.38	6.19	3.34	0.79	0.15	0.07	0.04	0.01
±0.8	0	±0.6	14.46	7.00	4.26	2.59	0.84	0.17	0.09	0.04	0.02
±0.8	0	±0.8	9.38	5.33	3.41	2.21	0.82	0.18	0.09	0.05	0.02
±0.8	±0.2	±0.2	34.37	14.11	7.24	3.74	0.76	0.14	0.07	0.03	0.01
±0.8	±0.2	±0.4	30.08	12.81	6.75	3.55	0.77	0.14	0.07	0.03	0.01
±0.8	±0.2	±0.6	18.04	8.76	5.10	2.98	0.80	0.17	0.09	0.04	0.02
±0.8	±0.2	±0.8	8.55	4.99	3.32	2.20	0.84	0.20	0.10	0.05	0.02
±0.8	±0.4	±0.4	26.61	11.53	6.23	3.37	0.76	0.15	0.08	0.04	0.02
±0.8	±0.4	±0.6	18.11	8.94	5.19	2.98	0.77	0.17	0.09	0.05	0.02
±0.8	±0.4	±0.8	9.32	5.31	3.51	2.32	0.81	0.21	0.11	0.06	0.03
±0.8	±0.6	±0.6	15.78	7.78	4.66	2.77	0.76	0.18	0.10	0.06	0.03
±0.8	±0.6	±0.8	8.98	5.27	3.46	2.25	0.78	0.22	0.13	0.08	0.04
±0.8	±0.8	±0.8	6.24	3.87	2.71	1.92	0.76	0.26	0.17	0.12	0.07
±0.8	±0.8	±1	1.00	1.00	1.00	1.00	1.00	0.37	0.24	0.17	0.10
±1 (sim)	0 (sim)	0 (sim)	50.72	20.04	10.00	4.99	1.00	0.20	0.10	0.05	0.02
±1 (calc)	0 (calc)	0 (calc)	50	20	10	5	1	0.2	0.1	0.05	0.02
±1	±0.2	±0.2	47.25	18.75	9.49	4.80	1.00	0.20	0.10	0.05	0.02
±1	±0.4	±0.4	38.57	16.16	8.45	4.42	1.00	0.23	0.12	0.06	0.03
±1	±0.6	±0.6	22.51	11.12	6.46	3.68	1.00	0.27	0.16	0.09	0.05
±1	±0.8	±0.8	9.30	5.65	3.83	2.60	1.00	0.38	0.26	0.18	0.11
±1	±1	±1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure A.1: Quantile-quantile plots, boxplots and regional association plots for 37 significant loci. This Figure consists of 37 Subfigures; one for each locus. **Quantile-quantile plot** The observed vs. expected distribution of $-\log_{10}$ p-values is plotted in the quantile-quantile plots. A deviation of the observed p-value distribution from the expected p-value distribution for small p-values (large $-\log_{10}$ p-values) indicates an association signal whereas a deviation for large p-values (small $-\log_{10}$ p-values) can indicate population stratification. The black line shows the results of the GWAS in KORA and the grey line of the GWAS in TwinsUK. **Boxplot** The measurements of the metabolic traits are stratified for the three genotypes (major allele homozygote, heterozygote, minor allele homozygote) at a SNP for KORA and TwinsUK separately. The number of samples per group is indicated above the plot. Notches indicate the 95 % confidence intervals around the means. The data is presented on a log-normal scale and normalised to the mean of the major allele homozygotes in each study. **Regional association plot** This plot shows the association signal for TwinsUK, KORA and the meta-analysis (Meta). Each point corresponds to a SNP in the region (genotyped SNPs are indicated in blue; imputed SNPs in black, the lead SNP in red). The genome-wide level of significance (2.0×10^{-12}) is indicated by horizontal grey lines. In the lower part of this plot are the genes (green arrows) summarised.

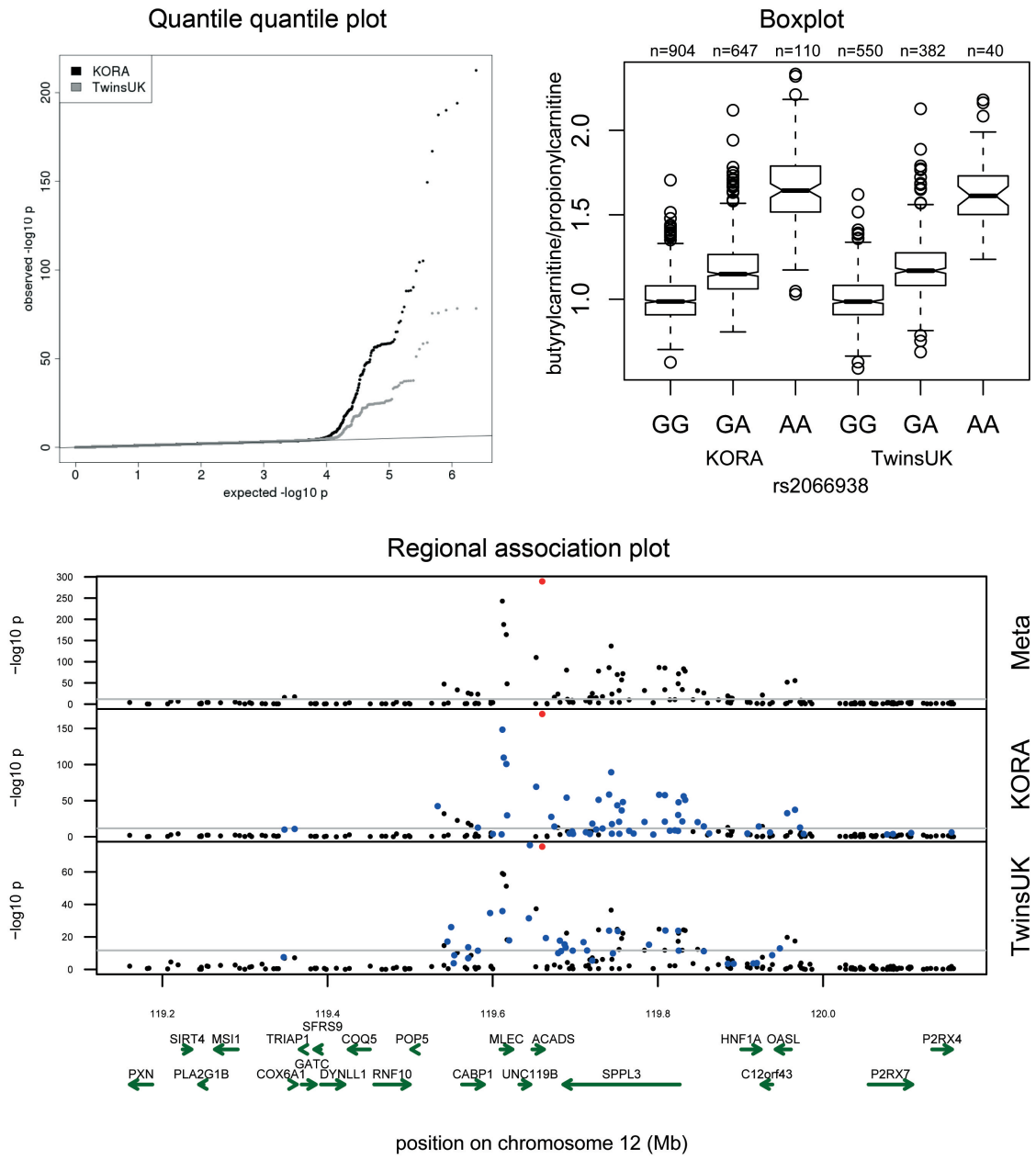

Figure A.1 (cont.)

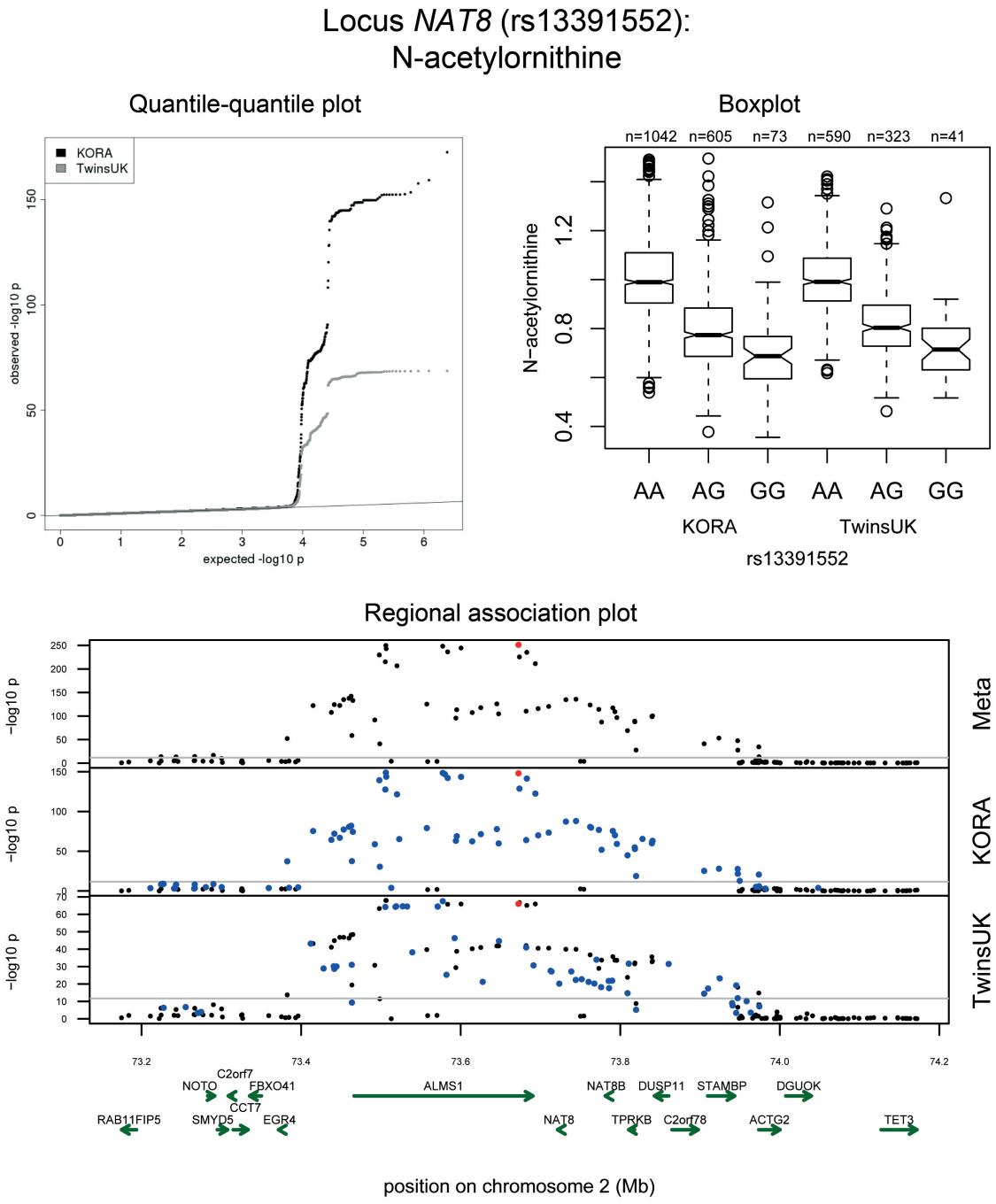

Figure A.1 (cont.)

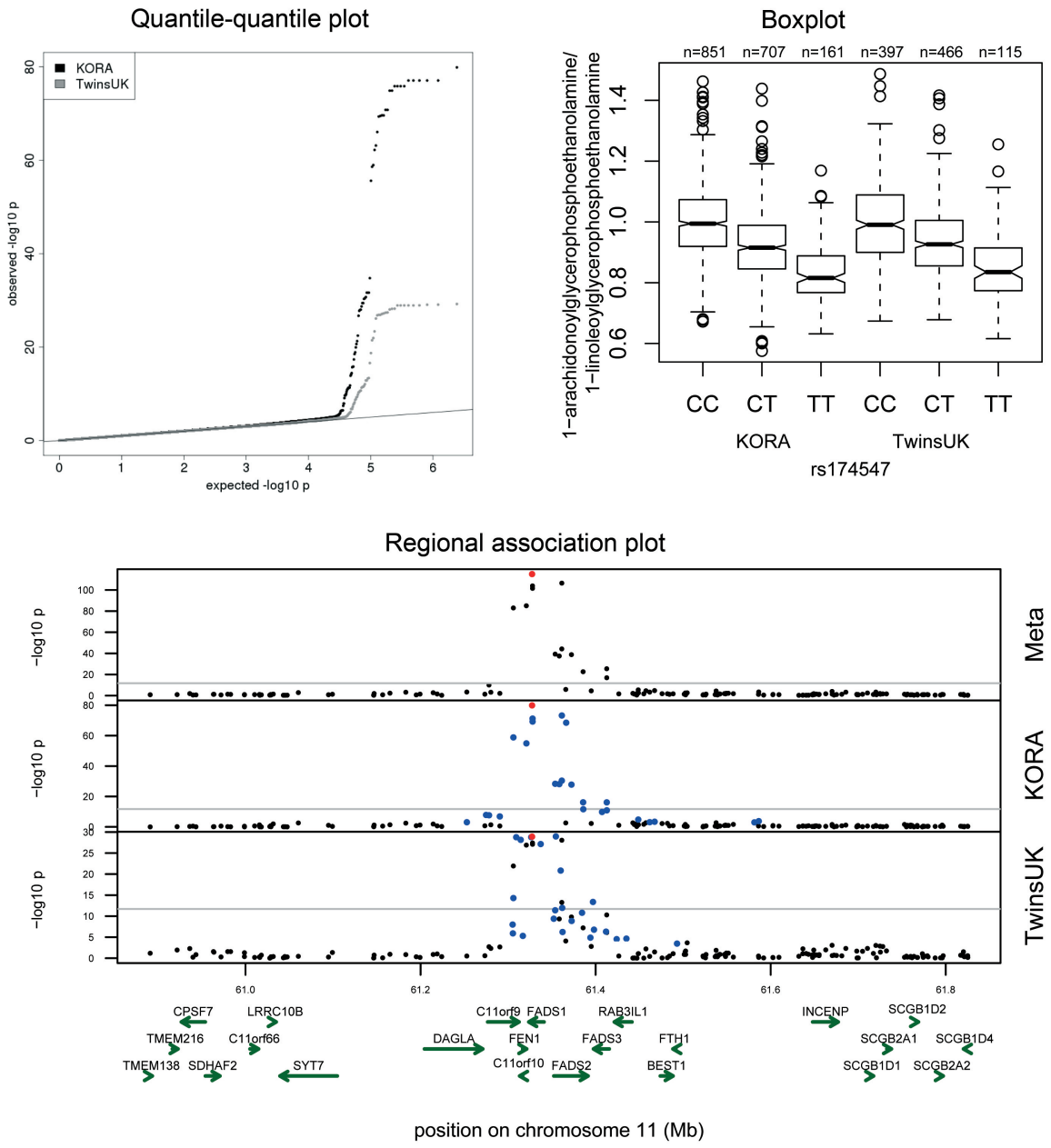

Figure A.1 (cont.)

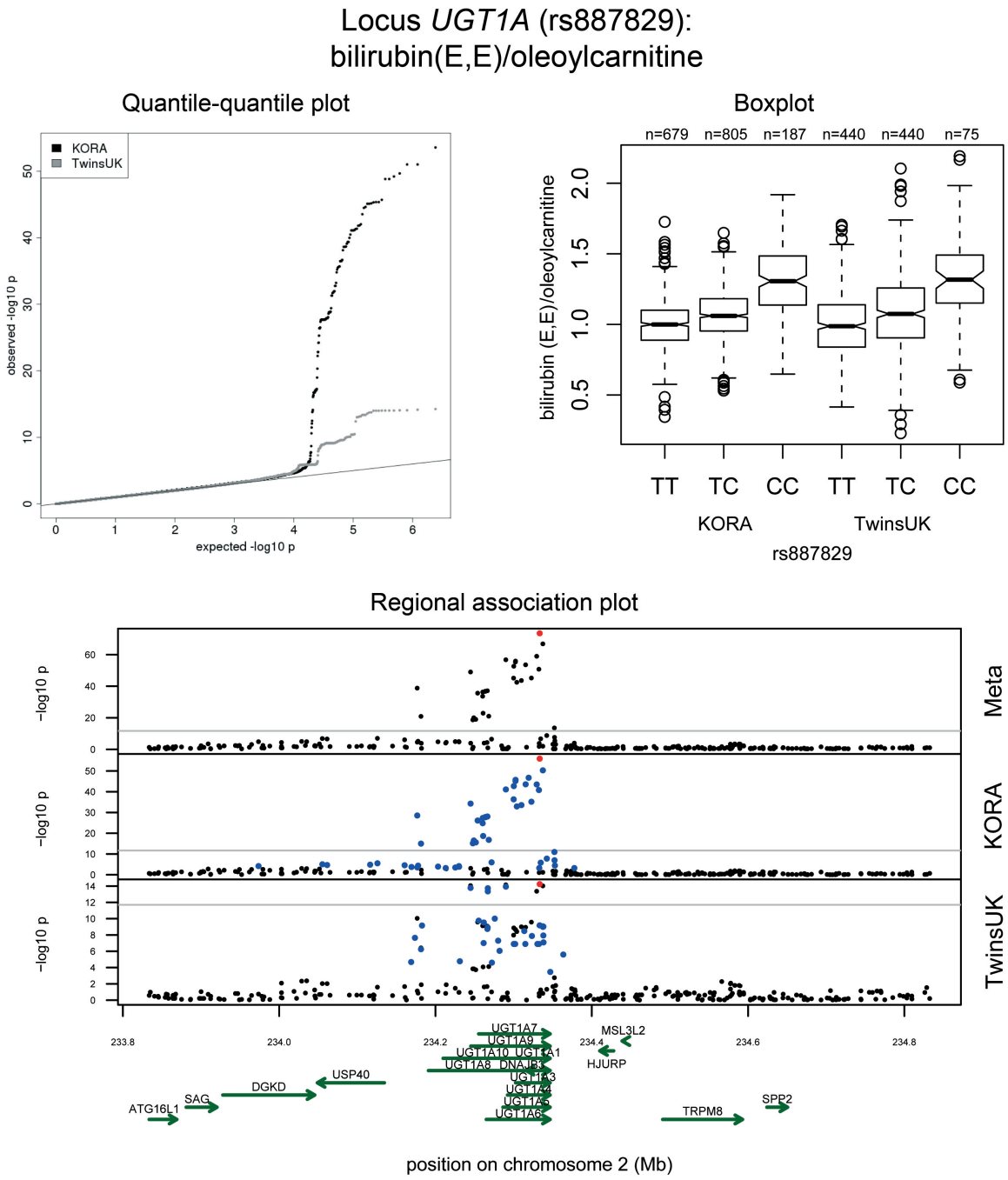

Figure A.1 (cont.)

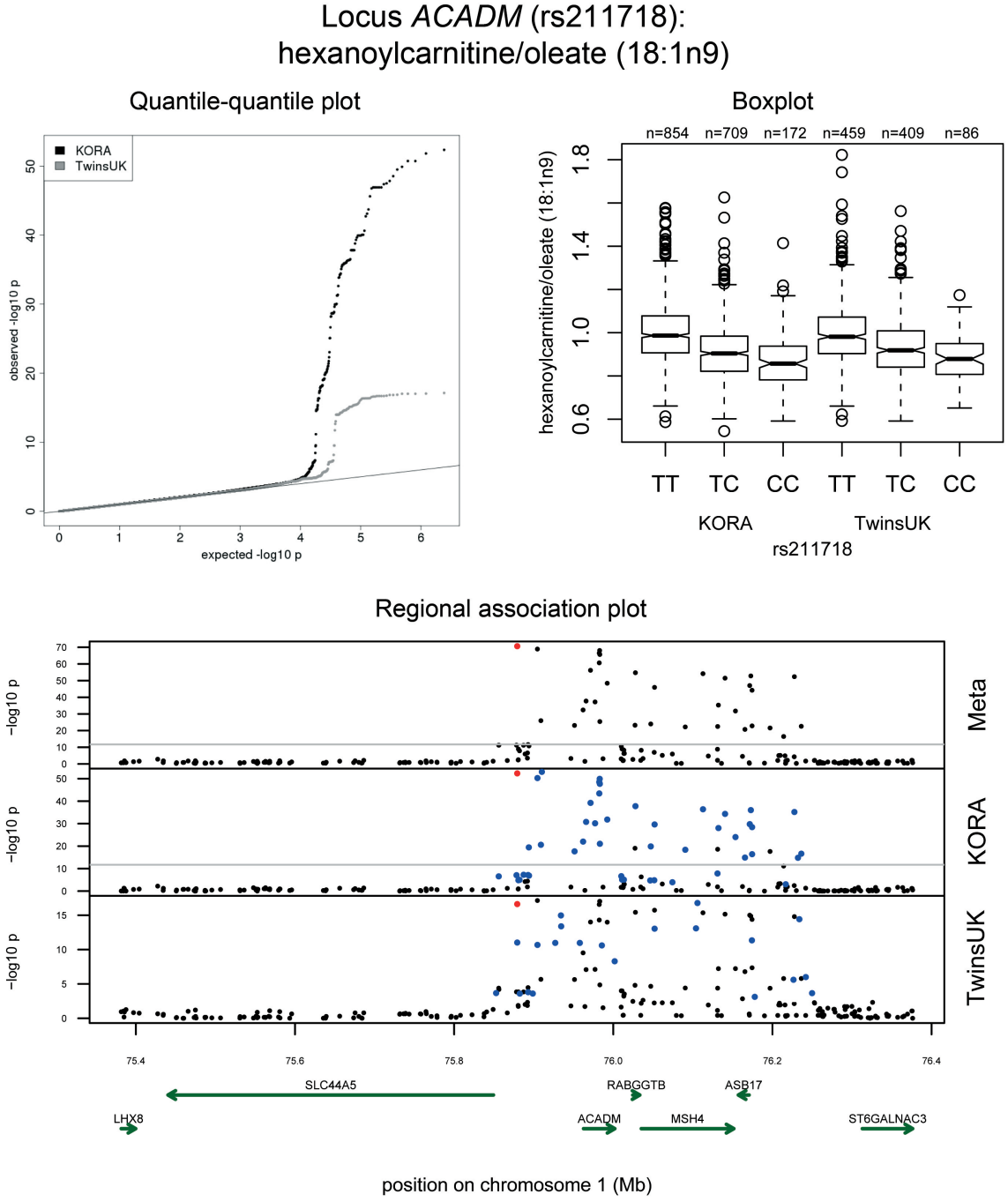

Figure A.1 (cont.)

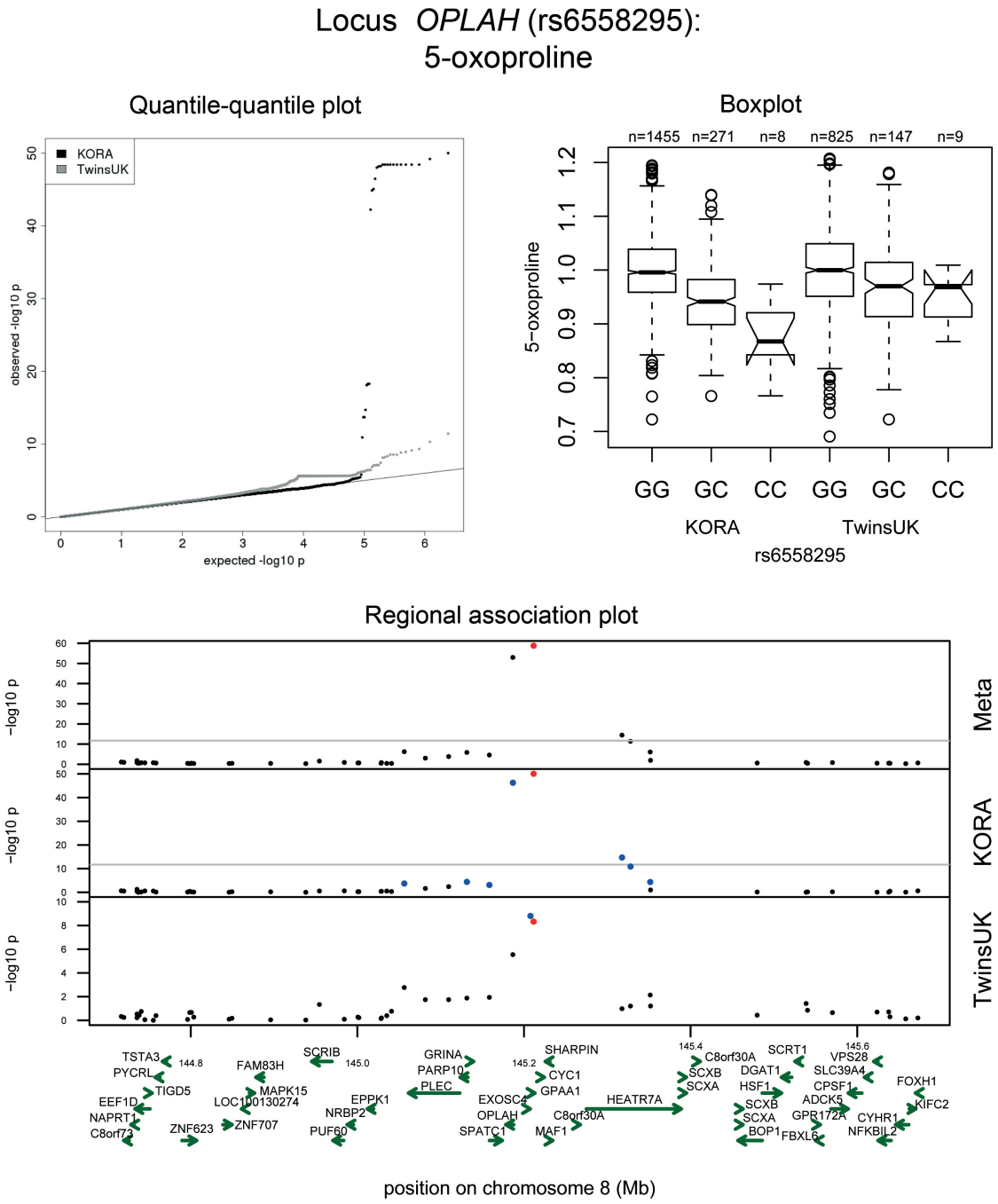

Figure A.1 (cont.)

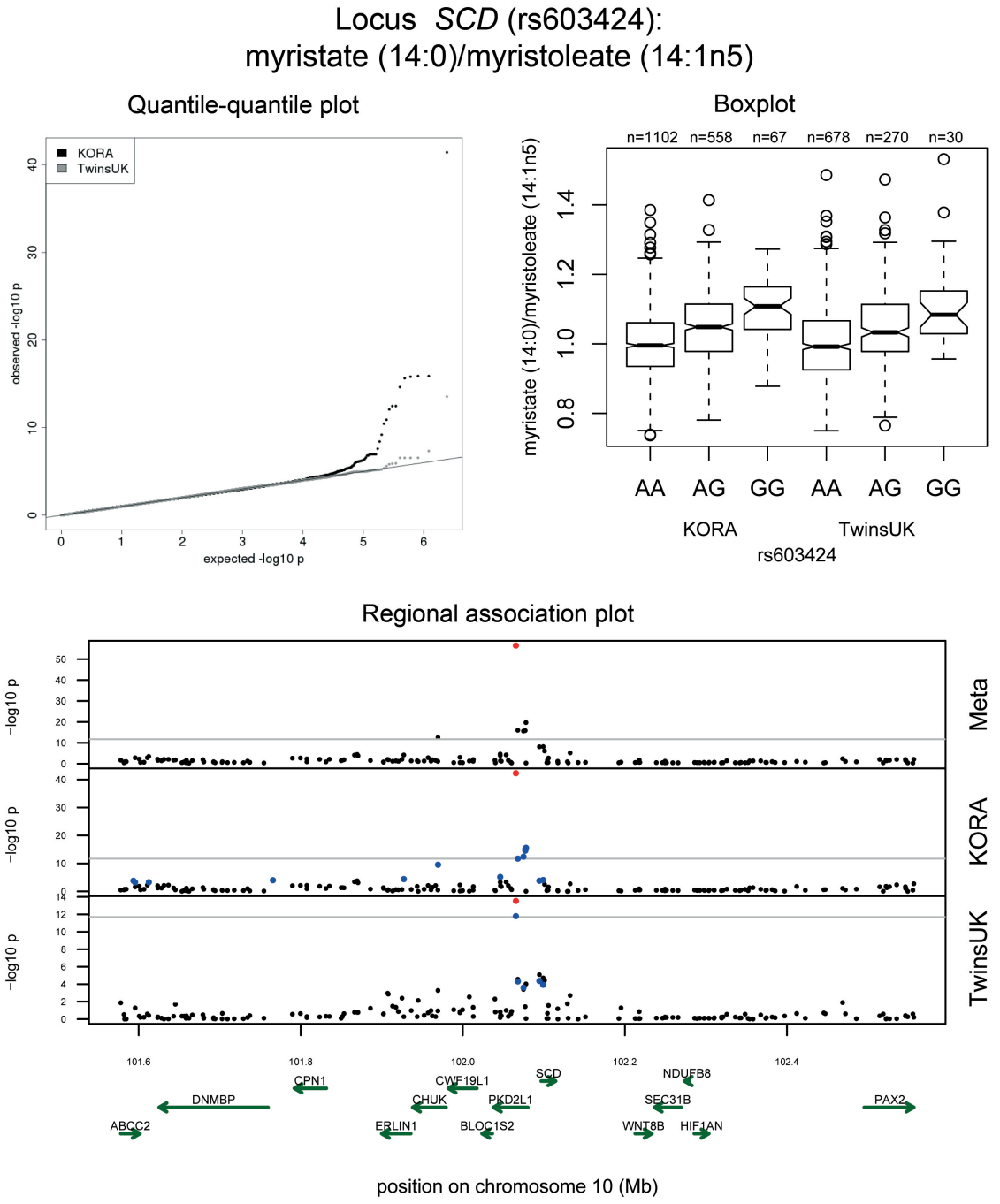

Figure A.1 (cont.)

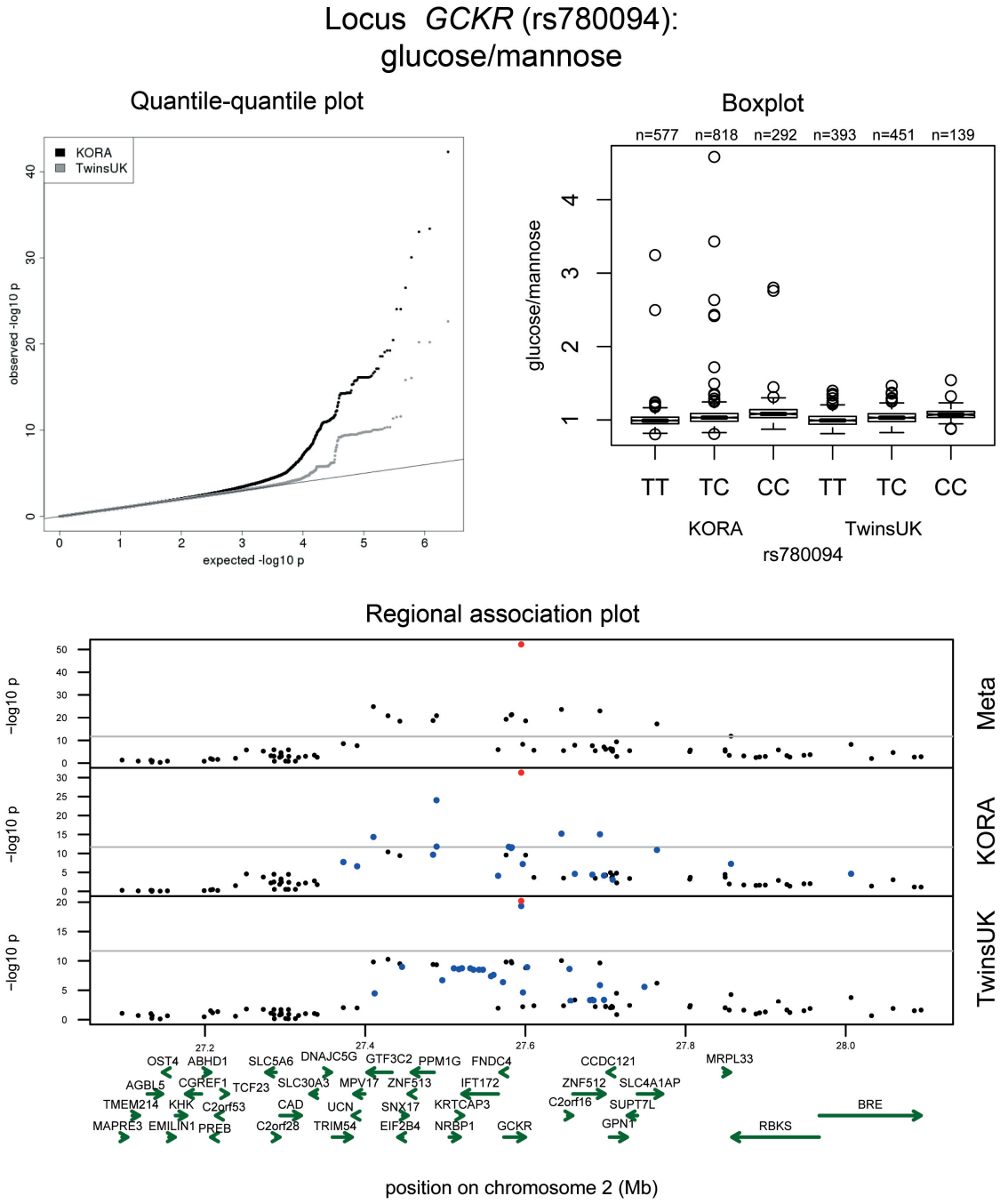

Figure A.1 (cont.)

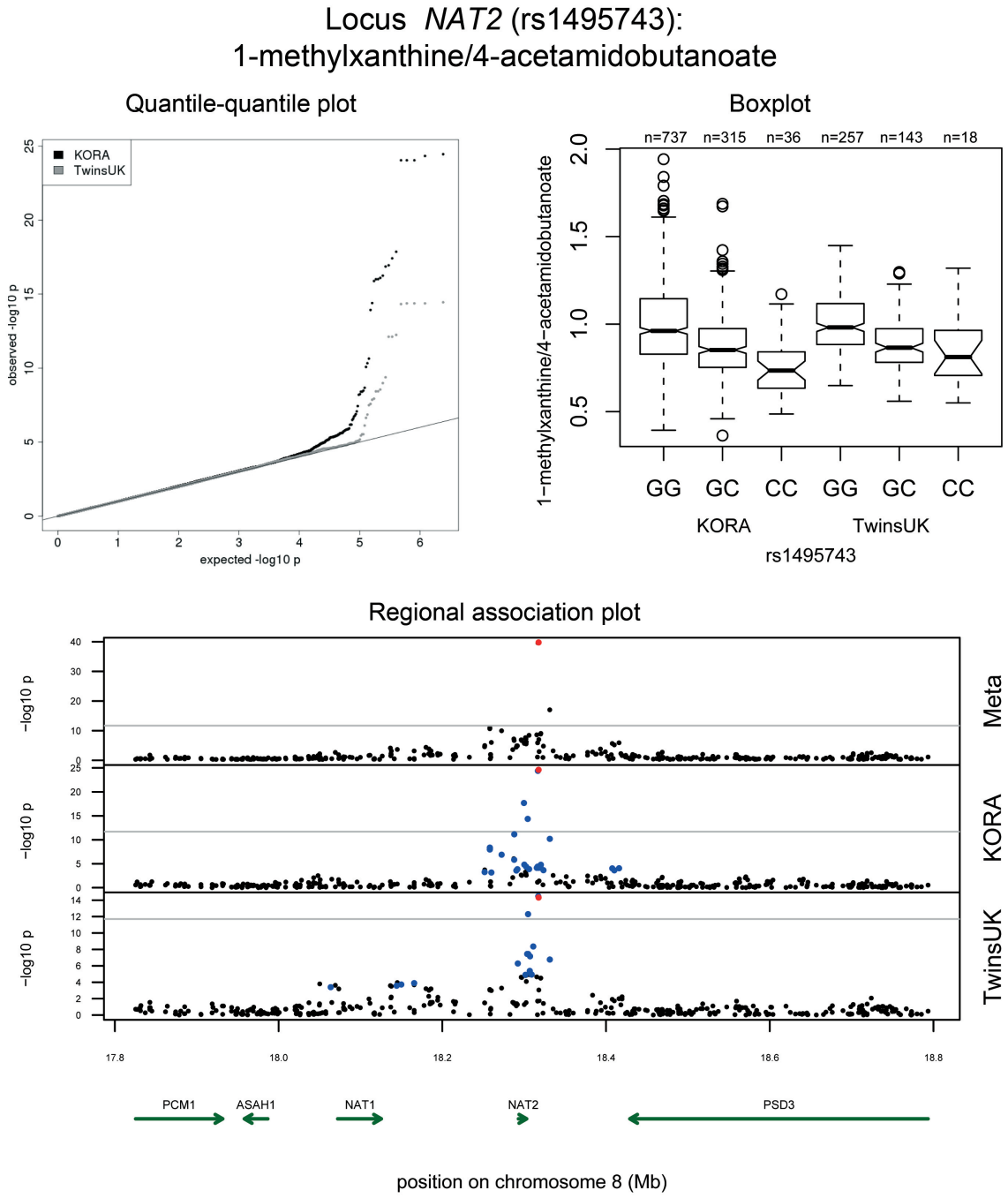

Figure A.1 (cont.)

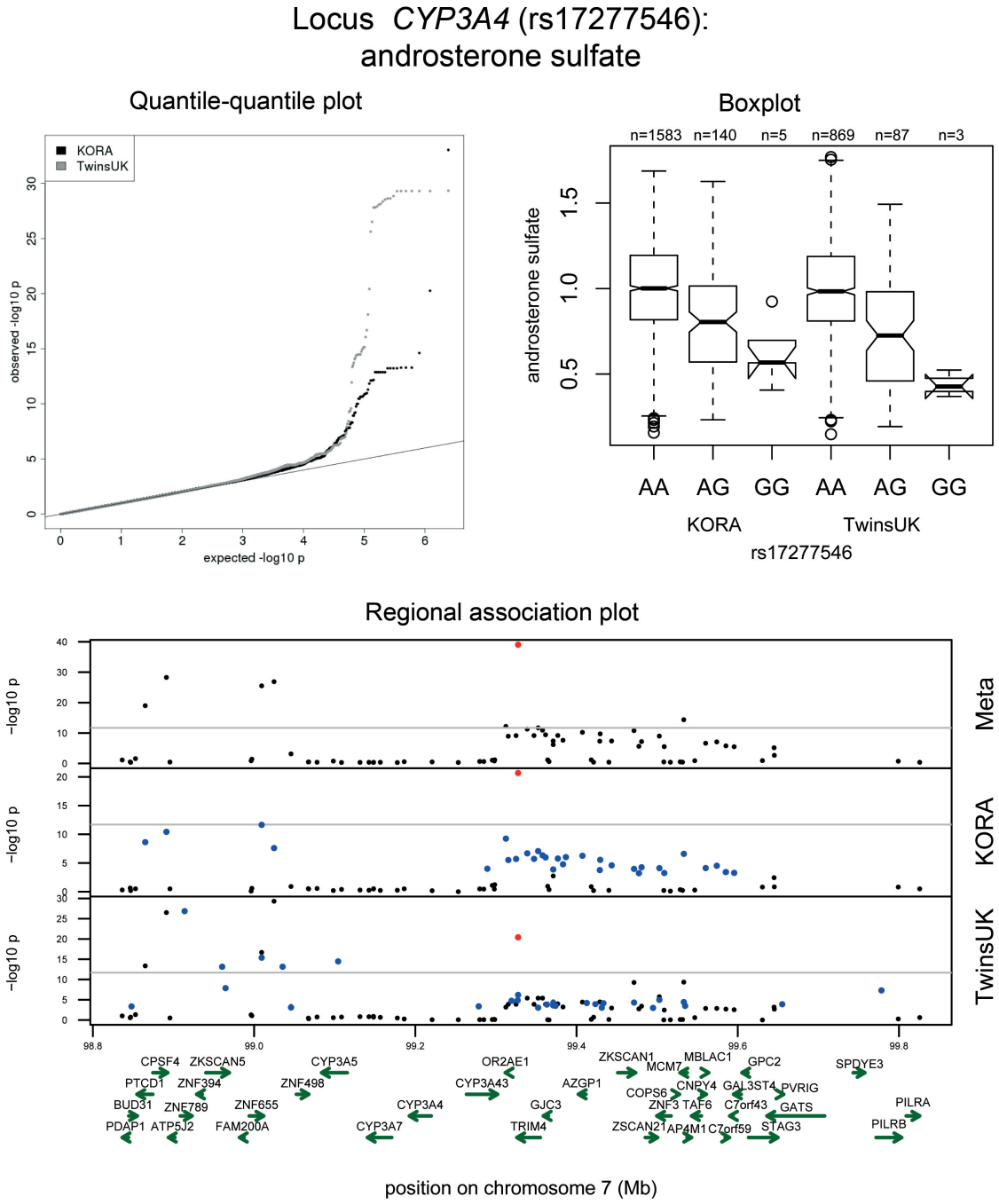

Figure A.1 (cont.)

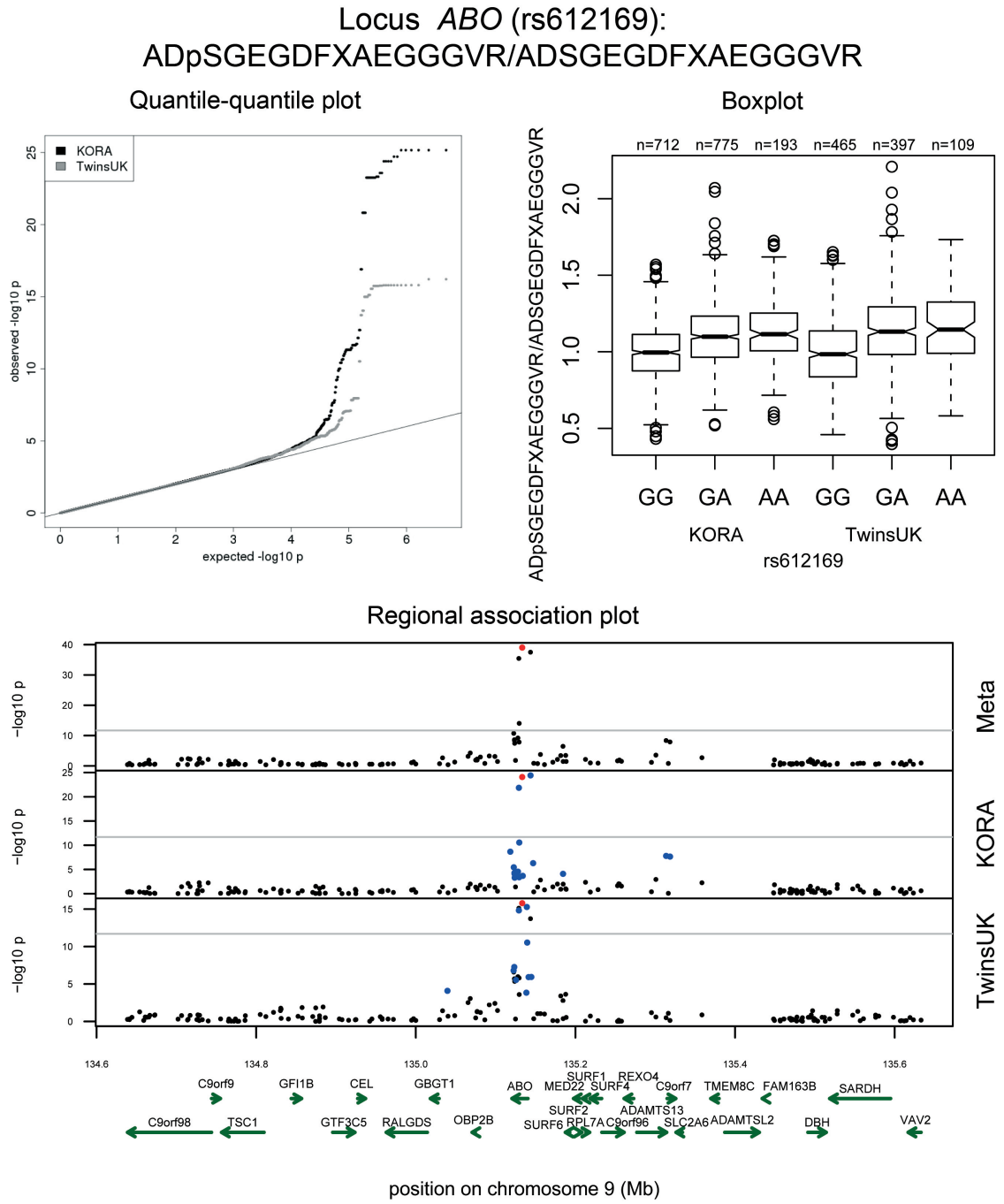

Figure A.1 (cont.)

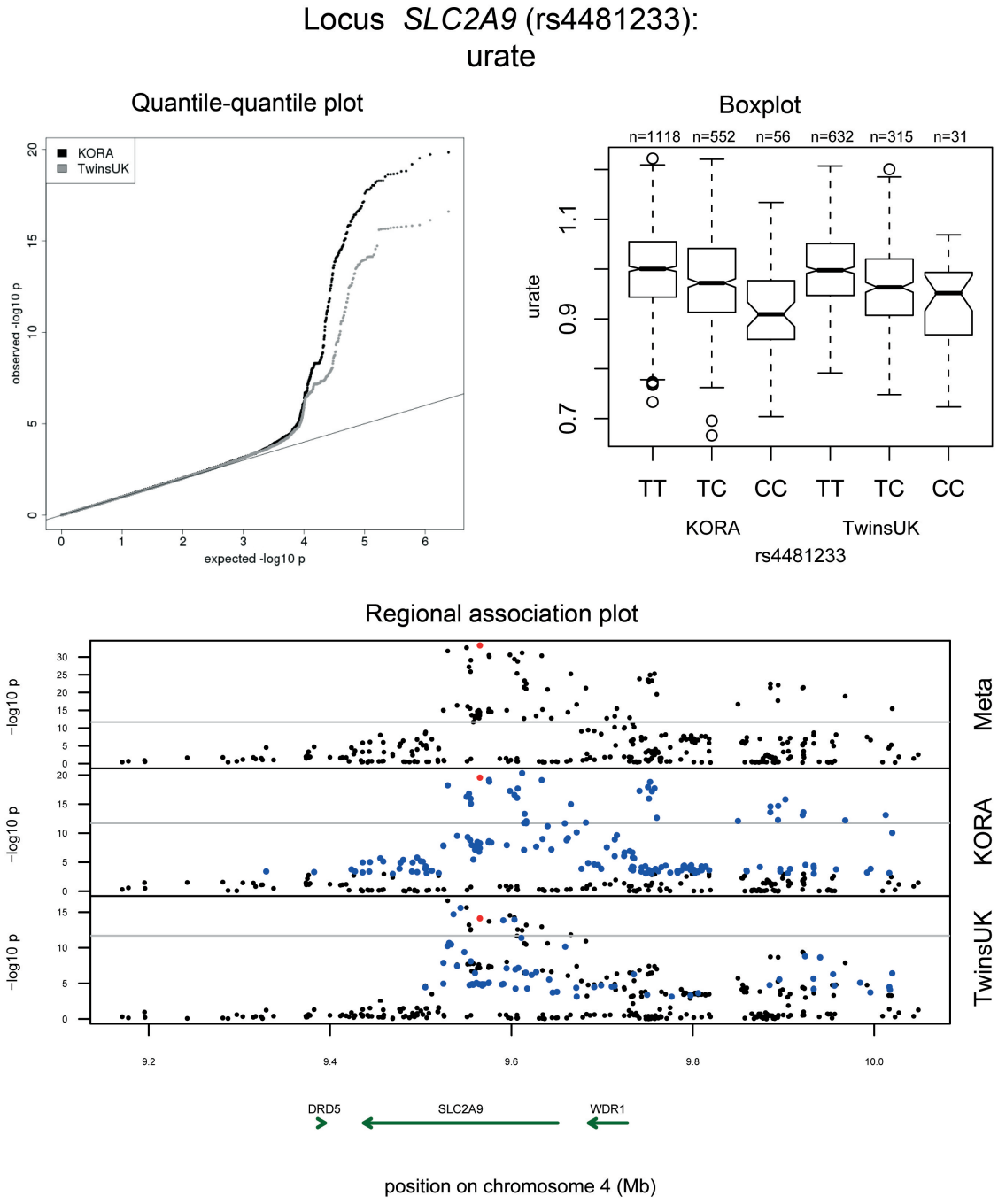

Figure A.1 (cont.)

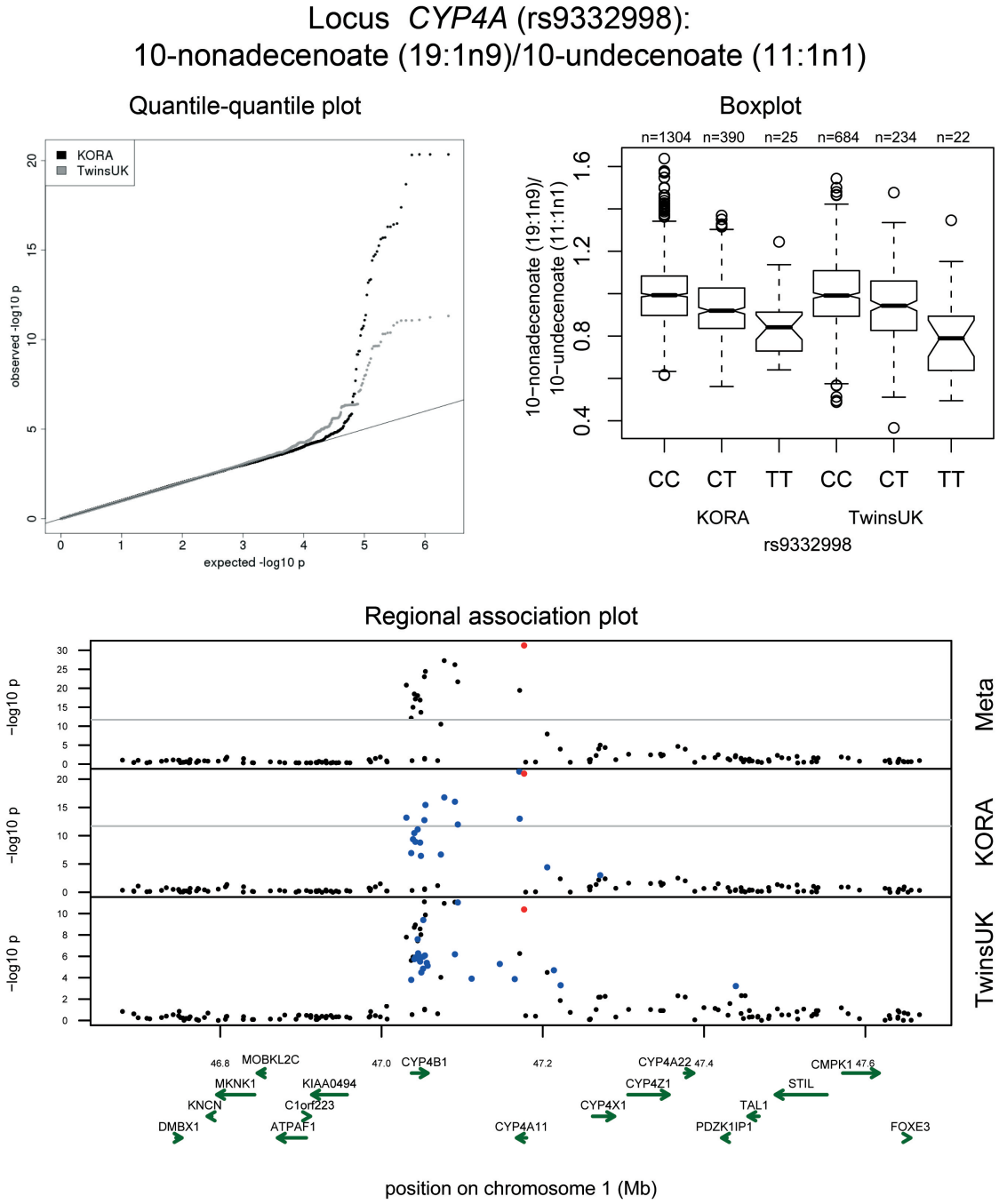

Figure A.1 (cont.)

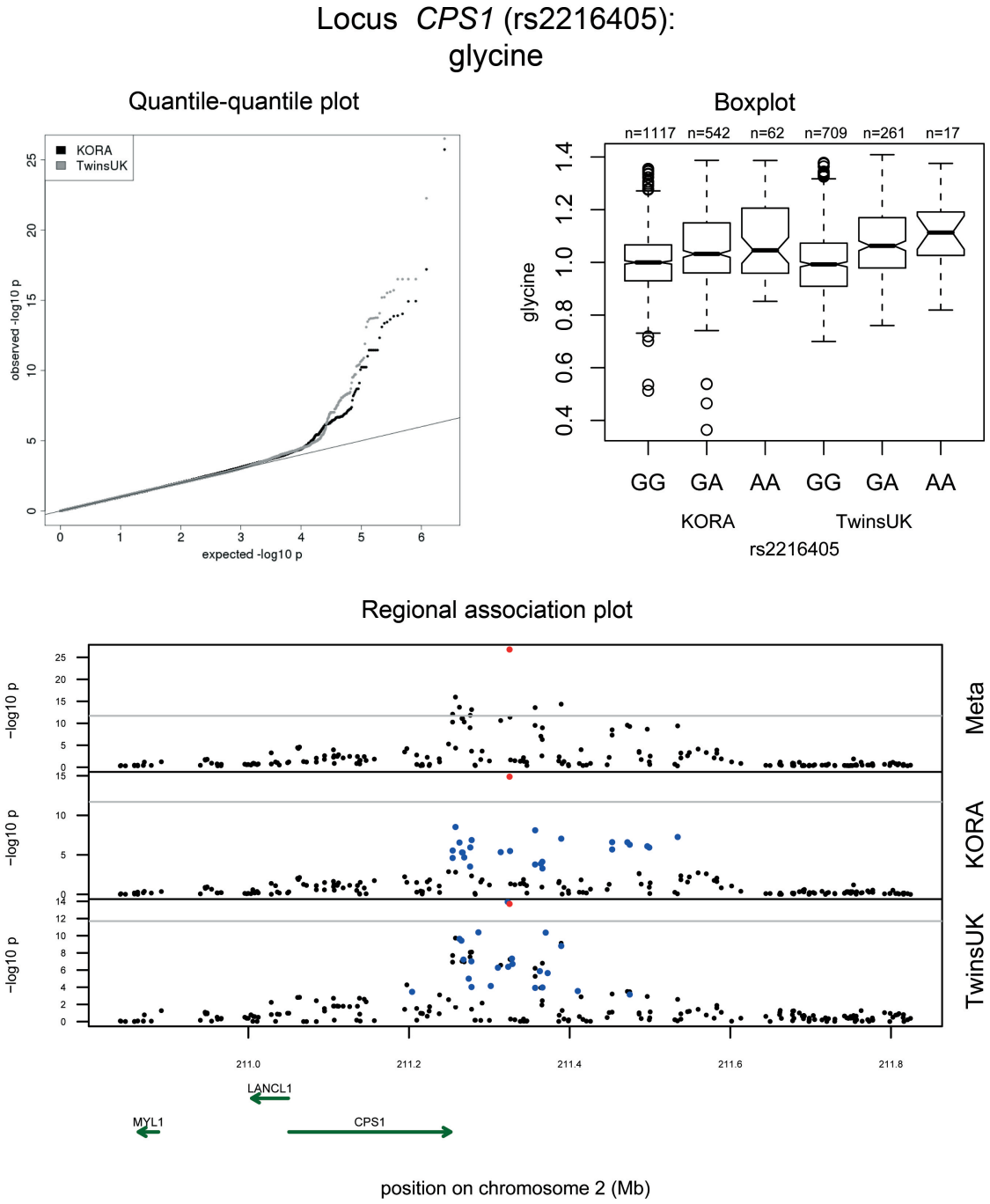

Figure A.1 (cont.)

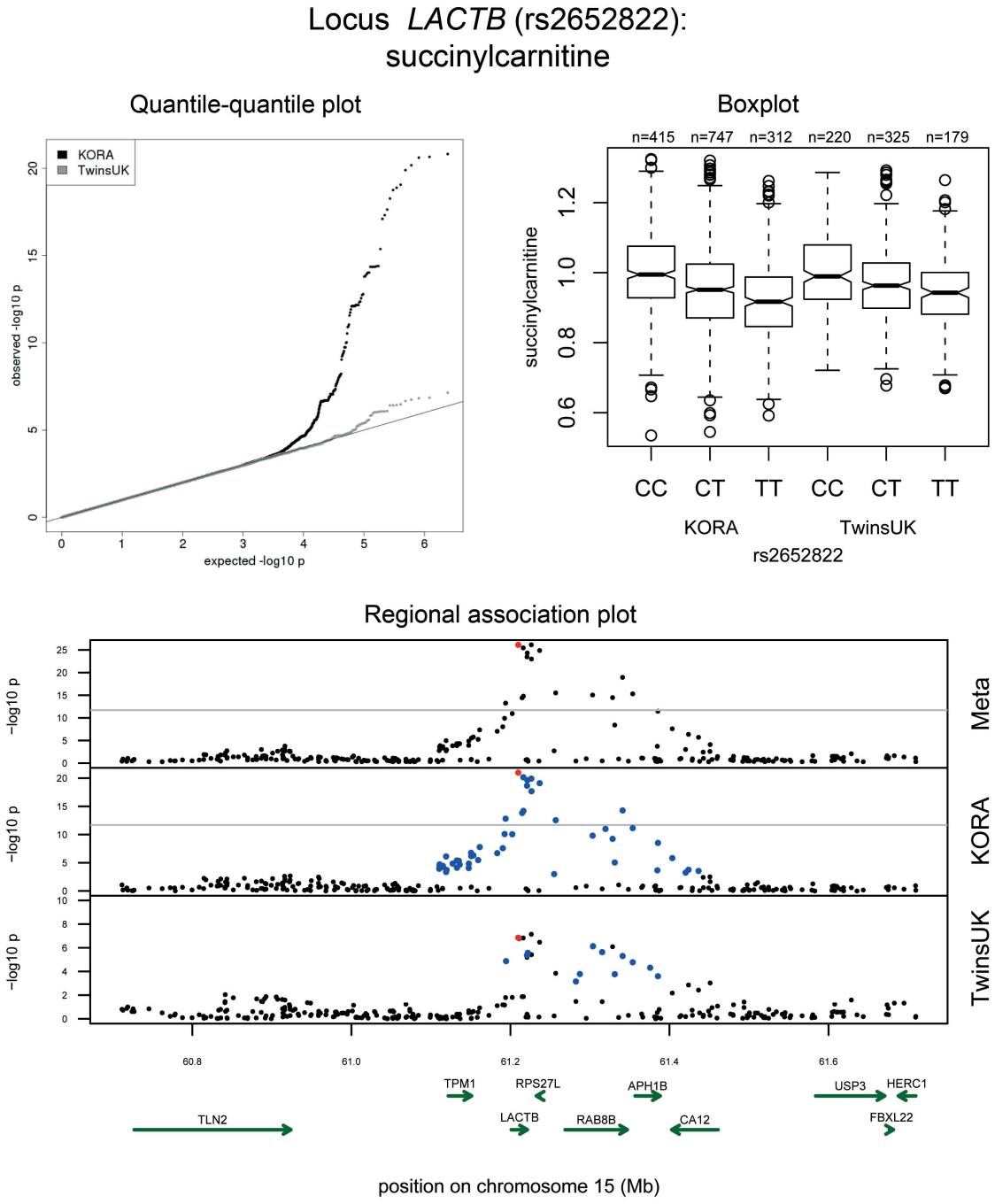

Figure A.1 (cont.)

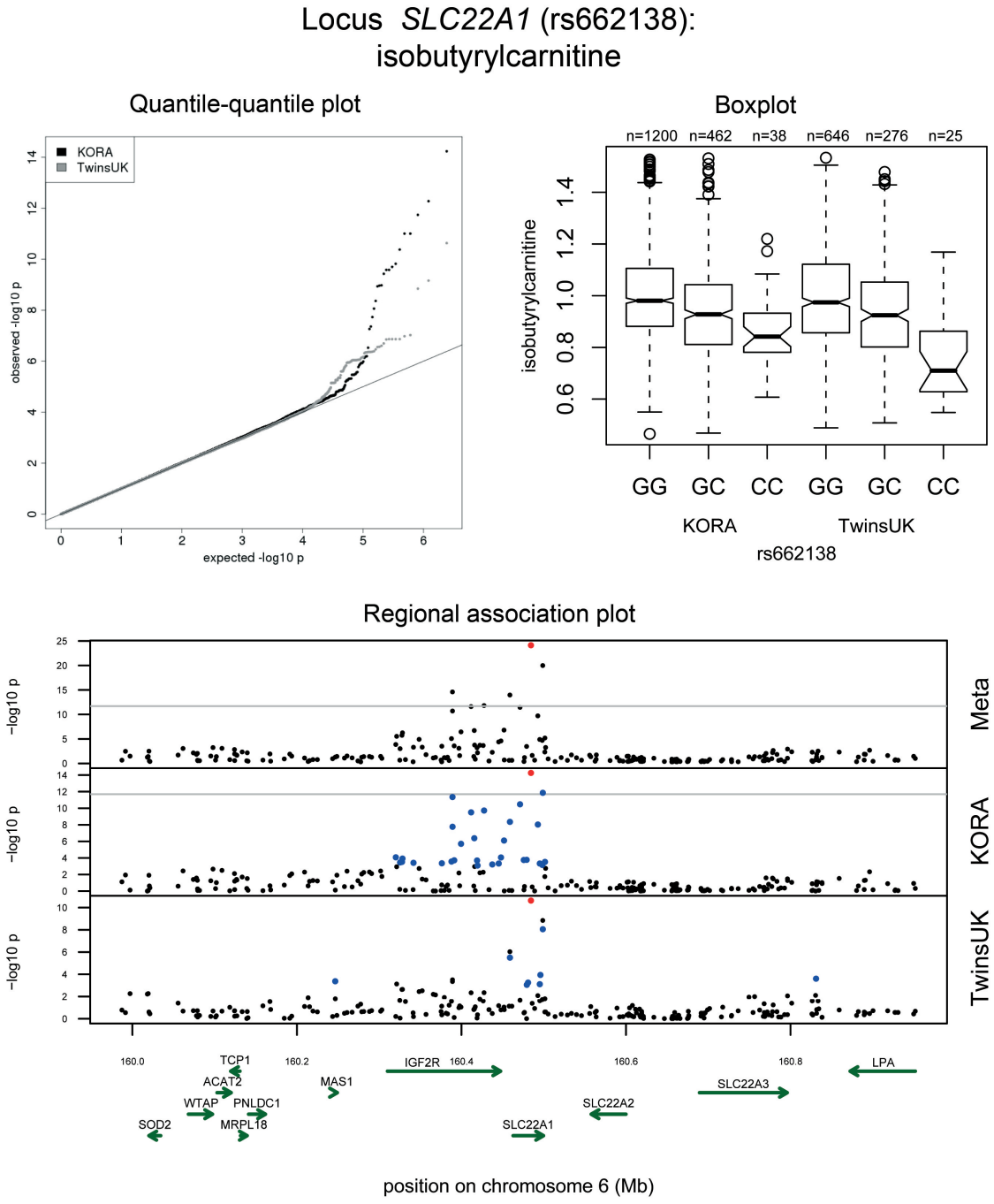

Figure A.1 (cont.)

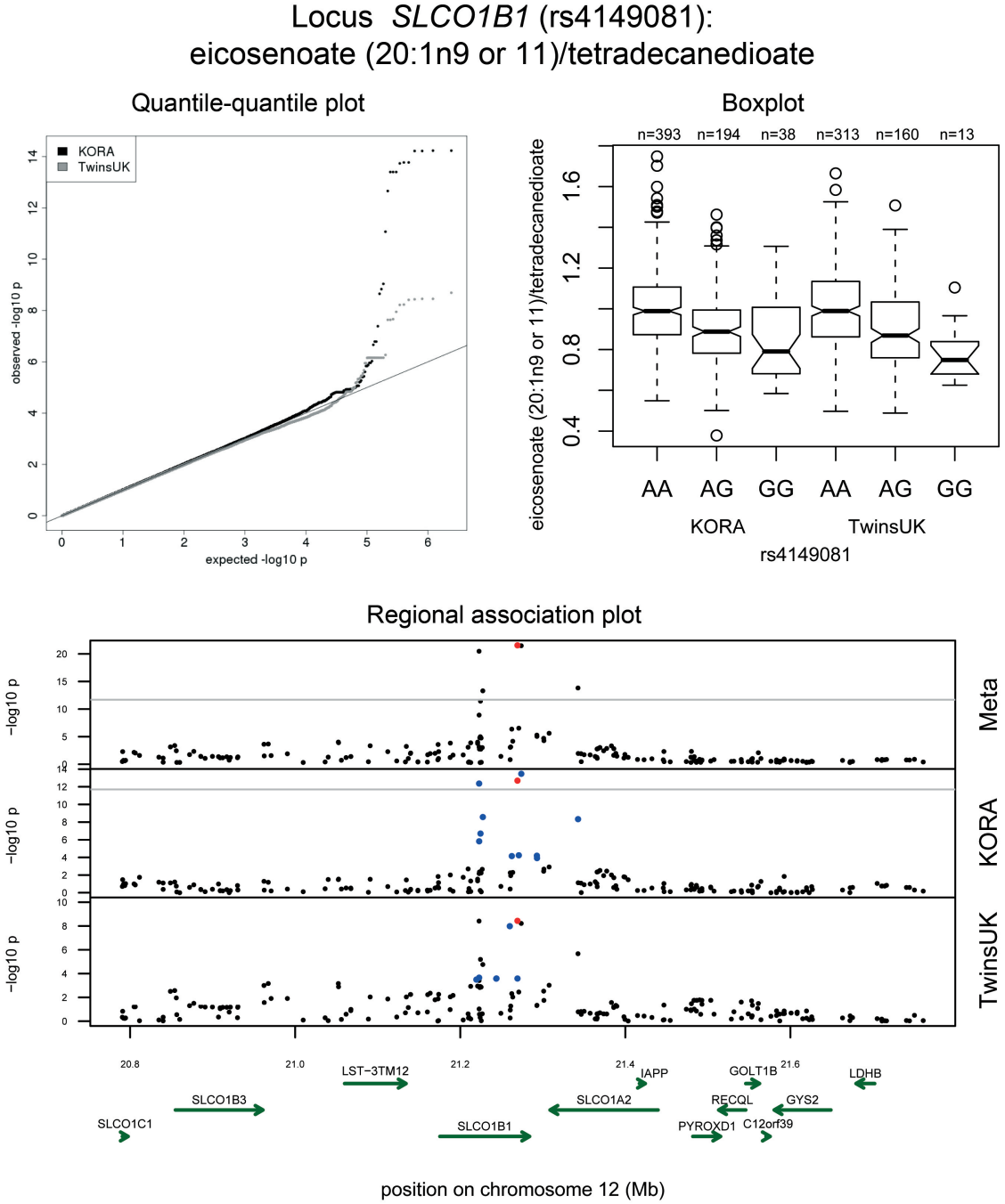

Figure A.1 (cont.)

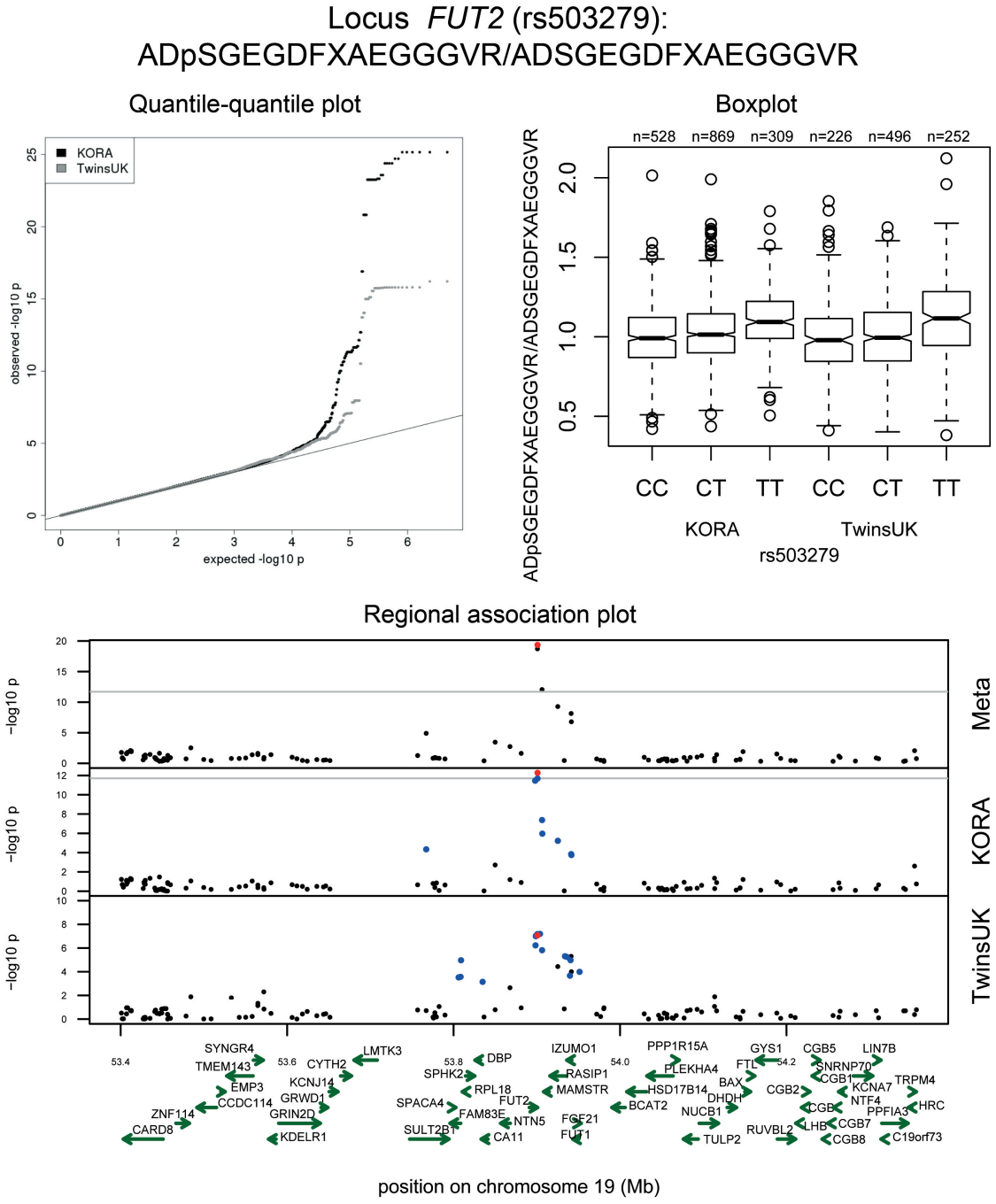

Figure A.1 (cont.)

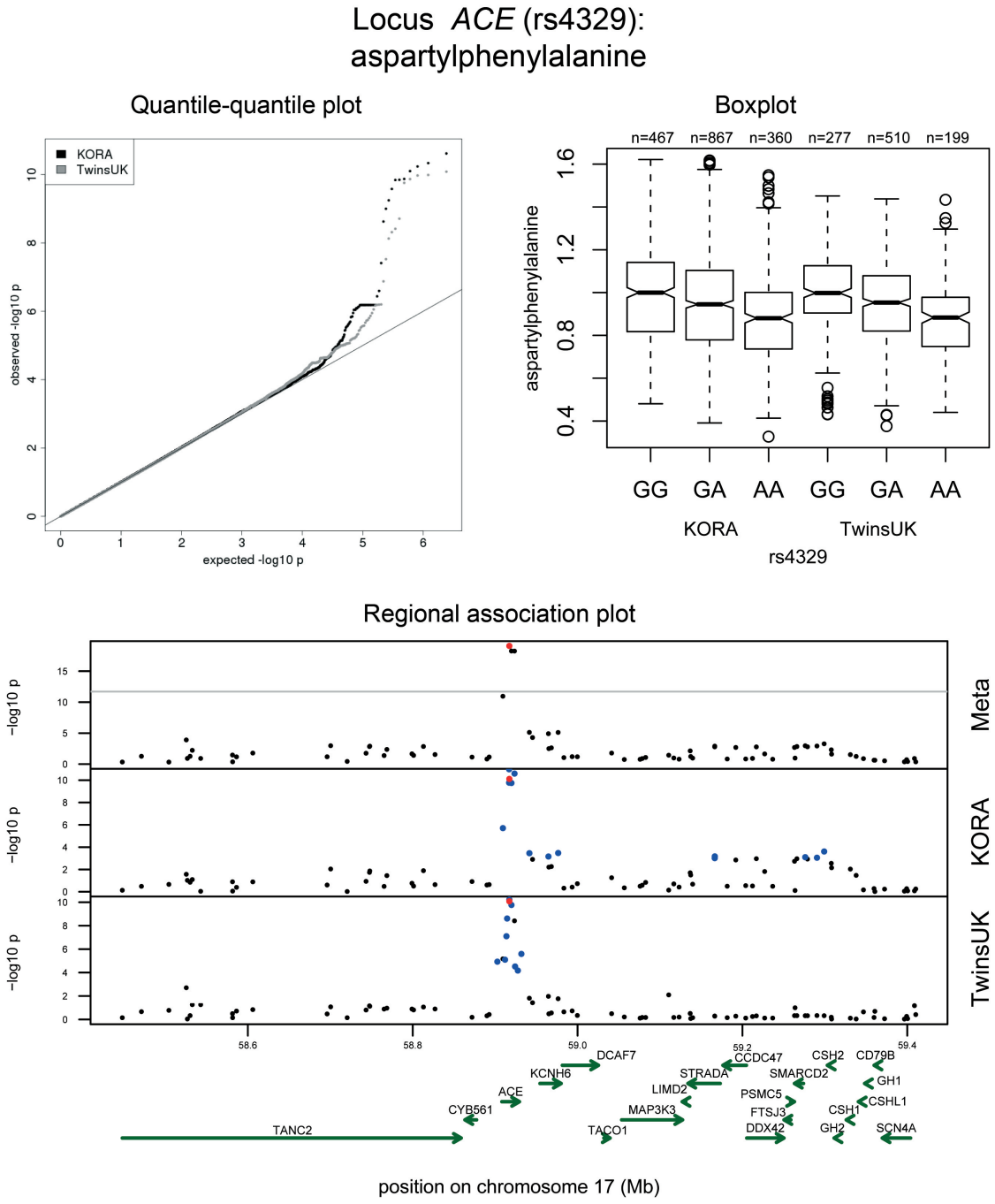

Figure A.1 (cont.)

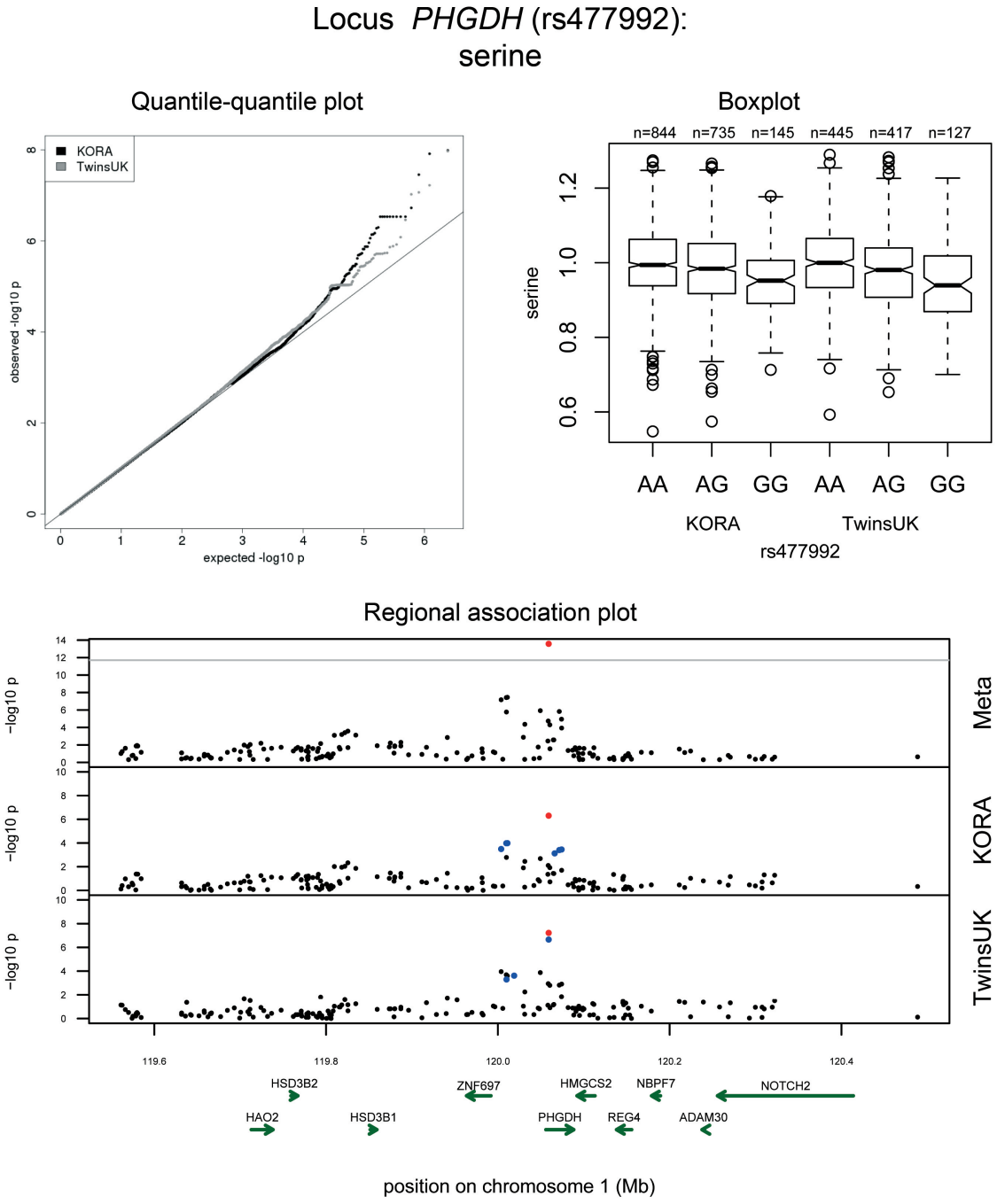

Figure A.1 (cont.)

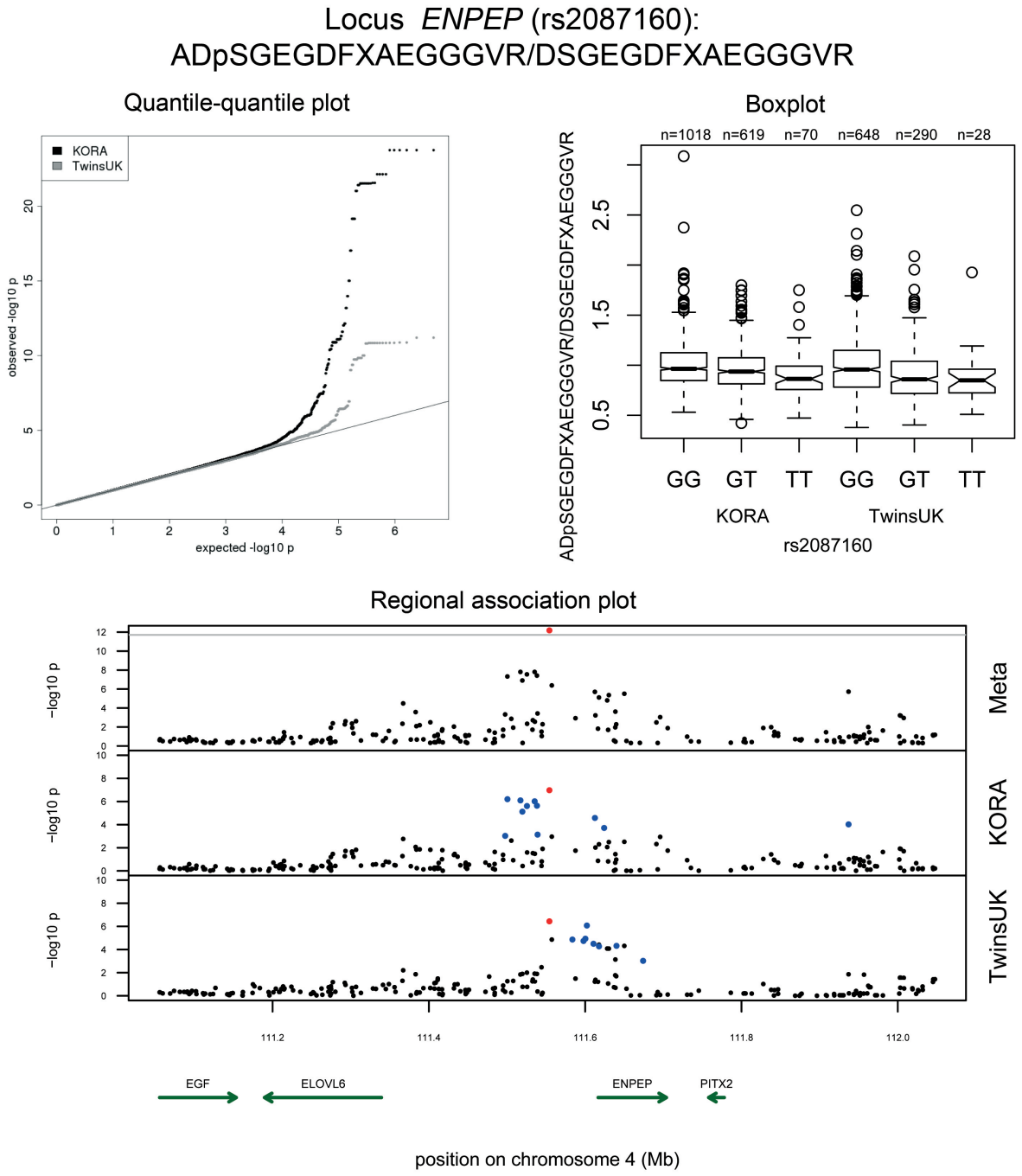

Figure A.1 (cont.)

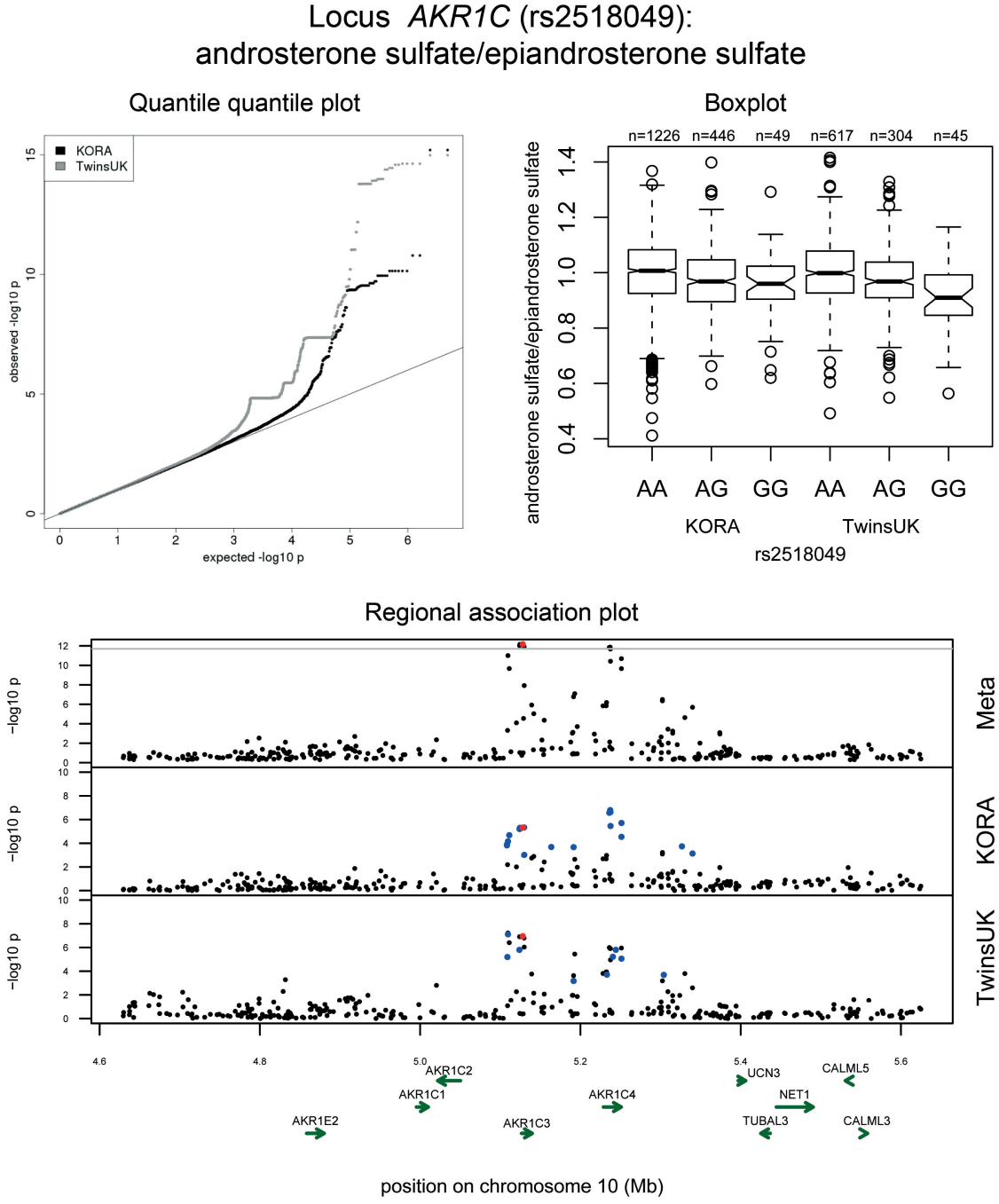

Figure A.1 (cont.)

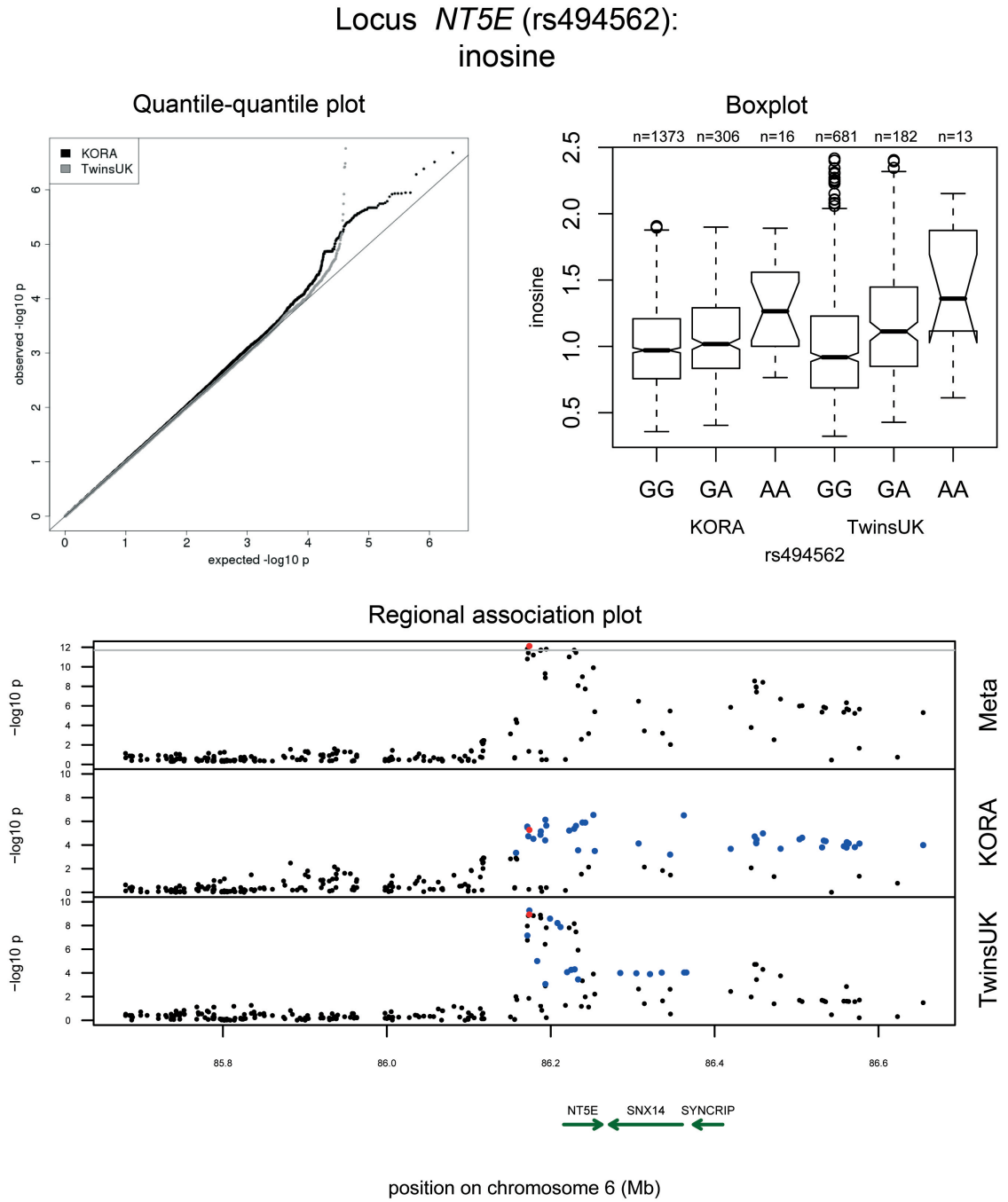

Figure A.1 (cont.)

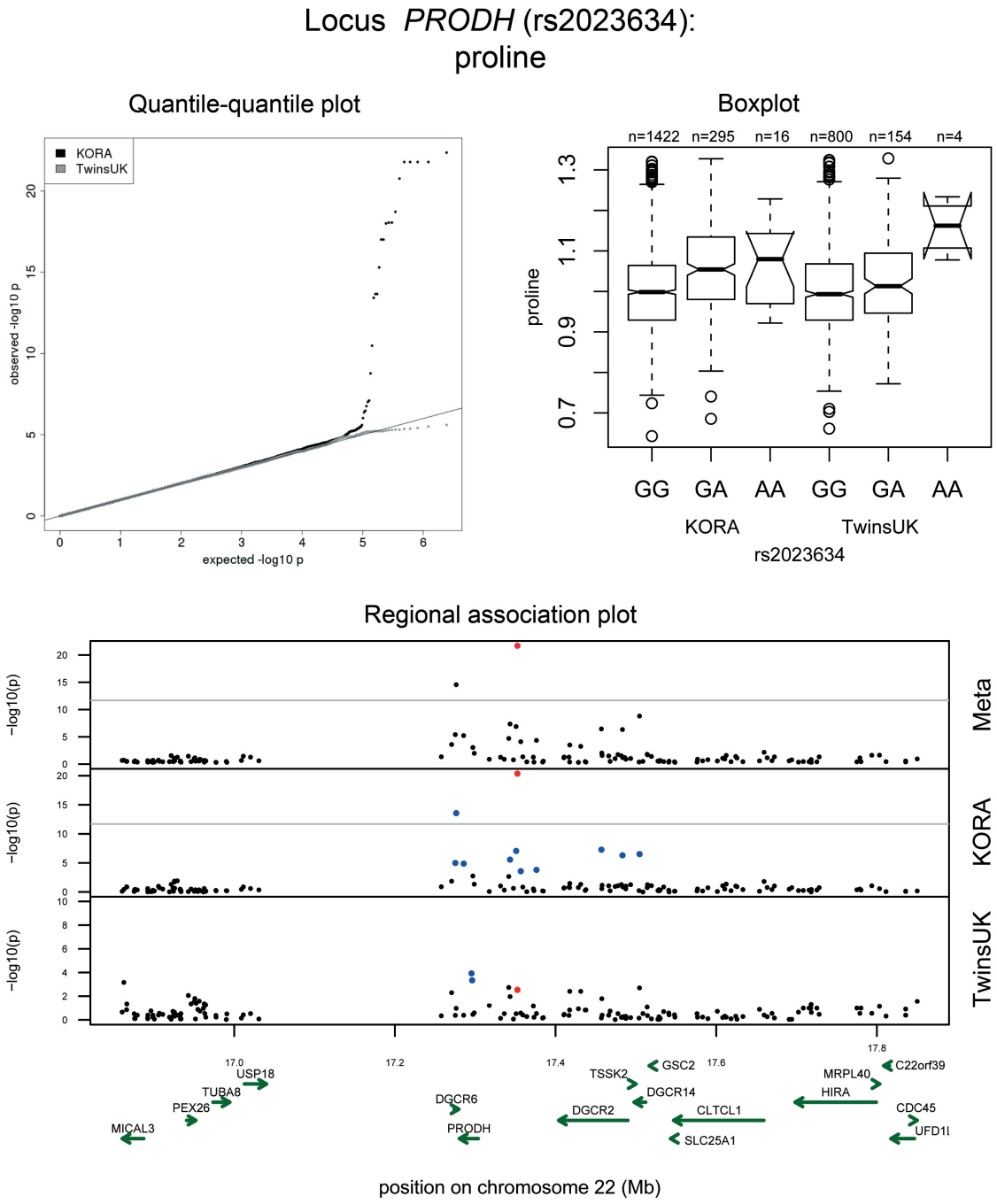

Figure A.1 (cont.)

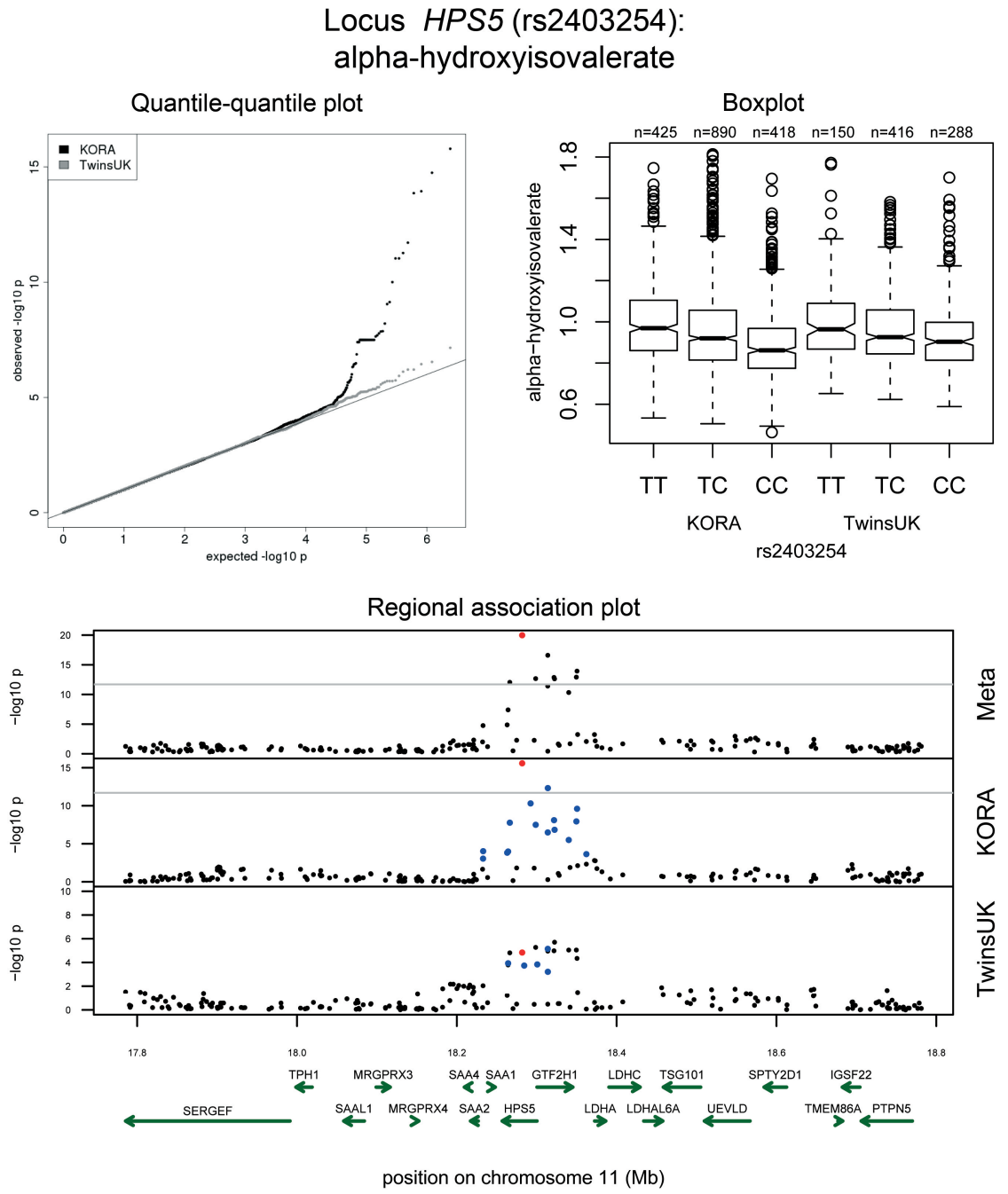

Figure A.1 (cont.)

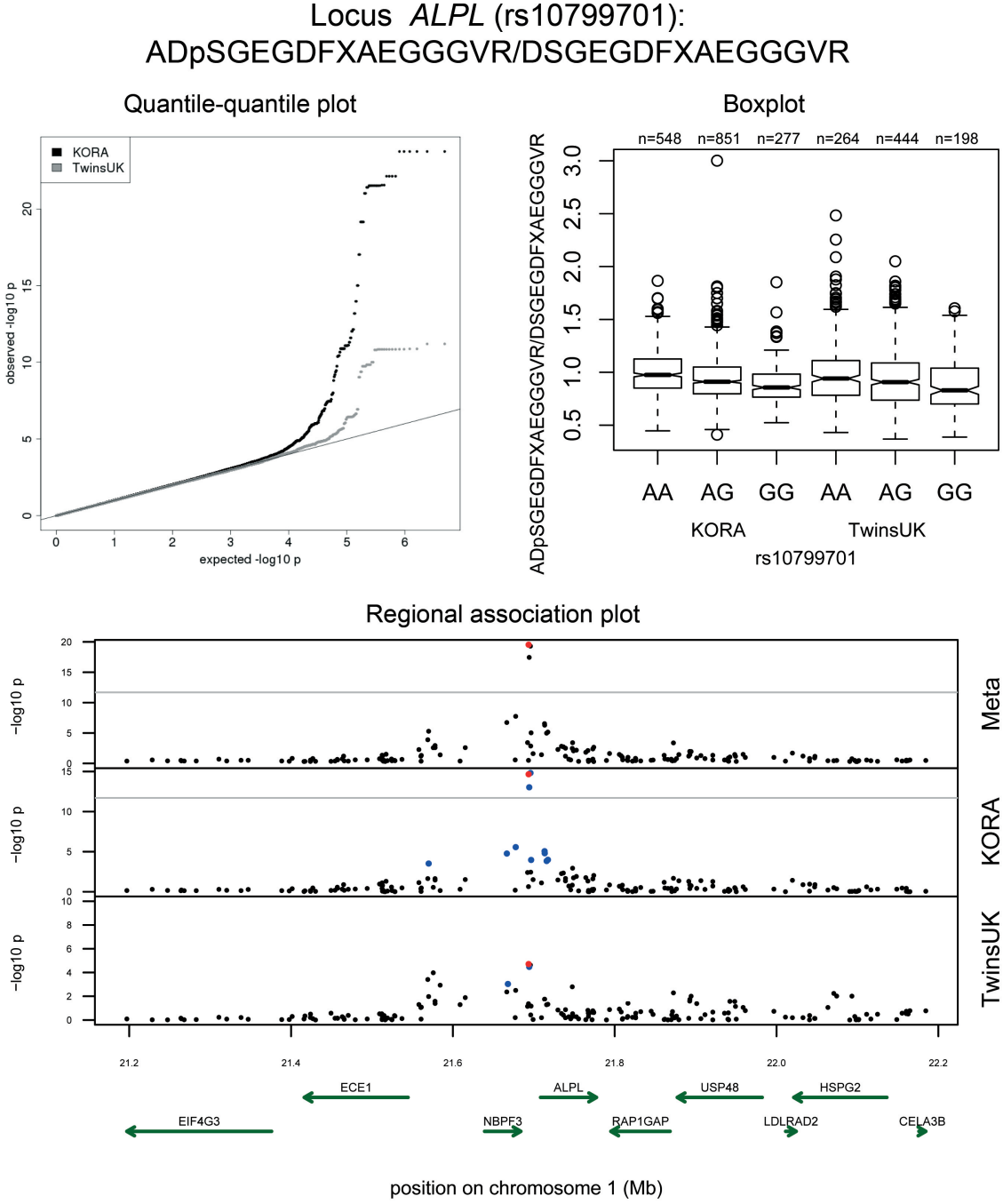

Figure A.1 (cont.)

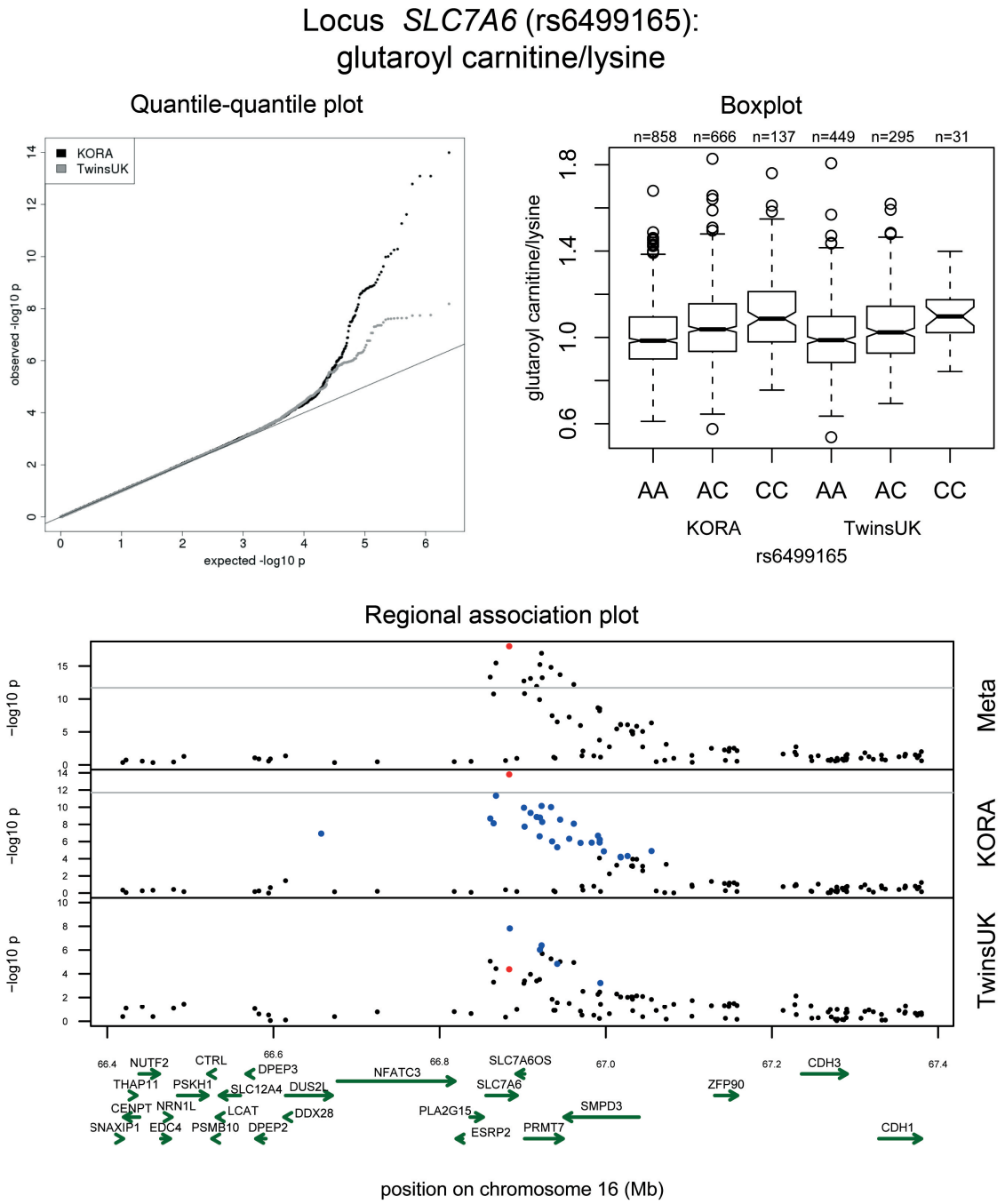

Figure A.1 (cont.)

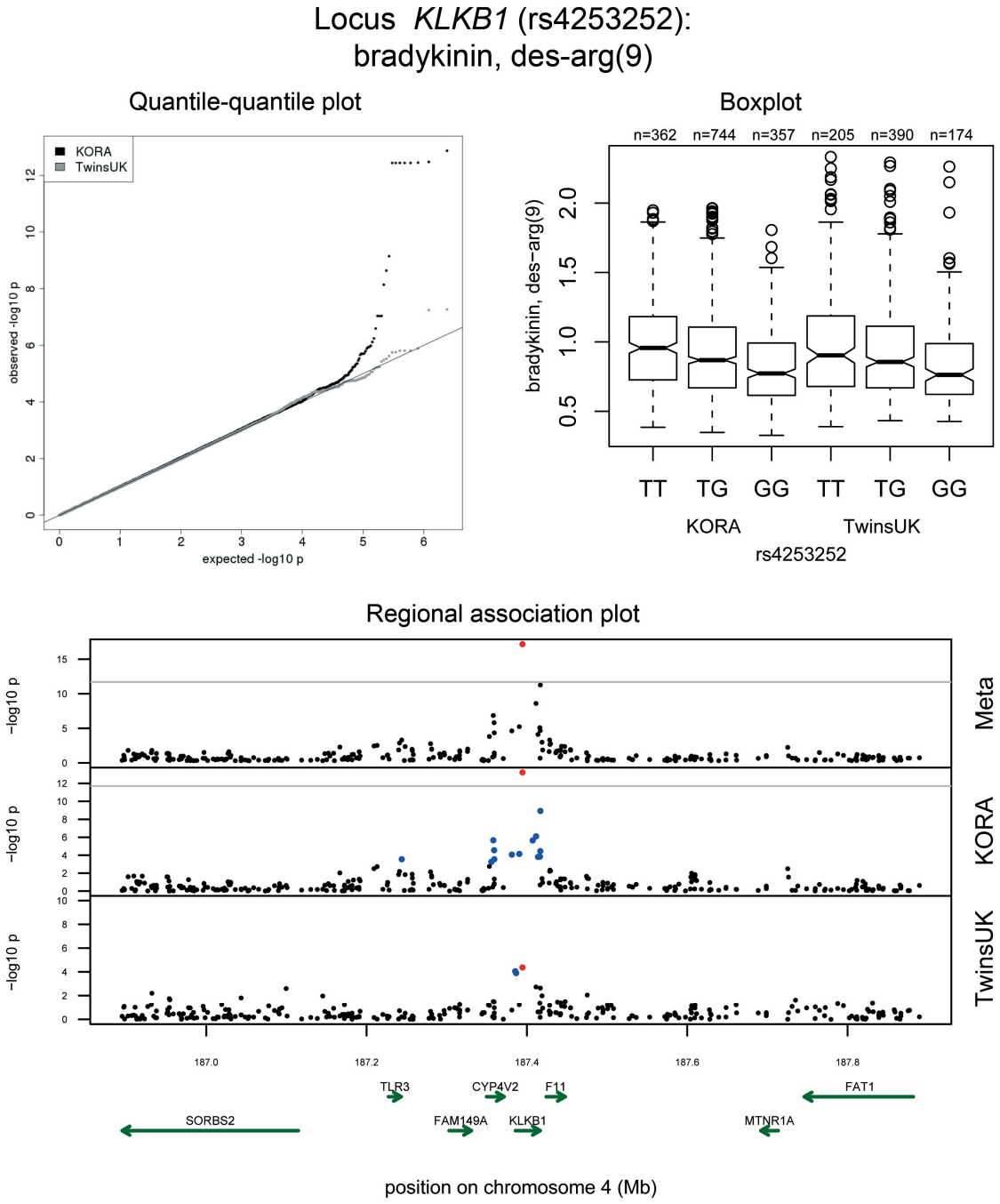

Figure A.1 (cont.)

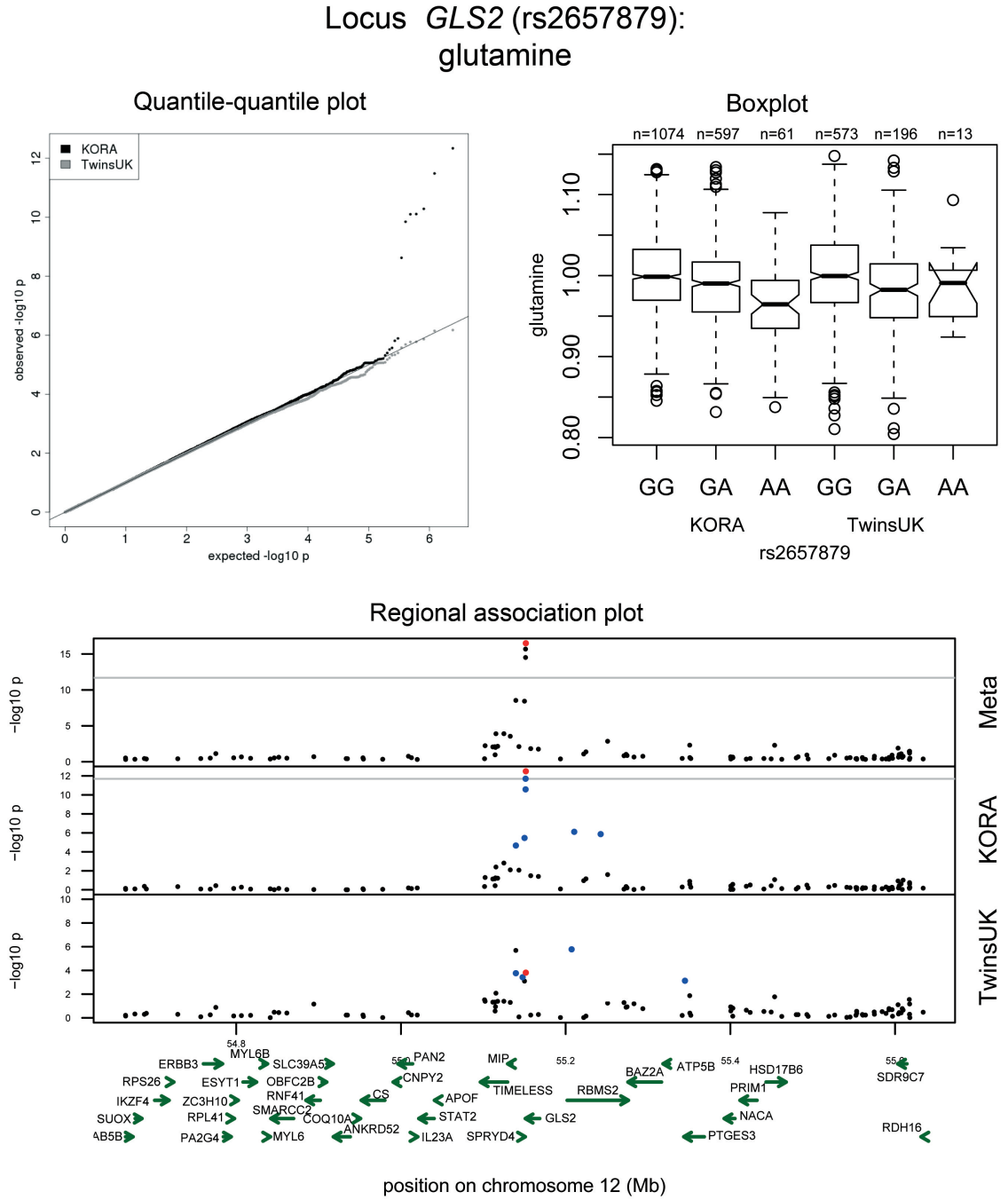

Figure A.1 (cont.)

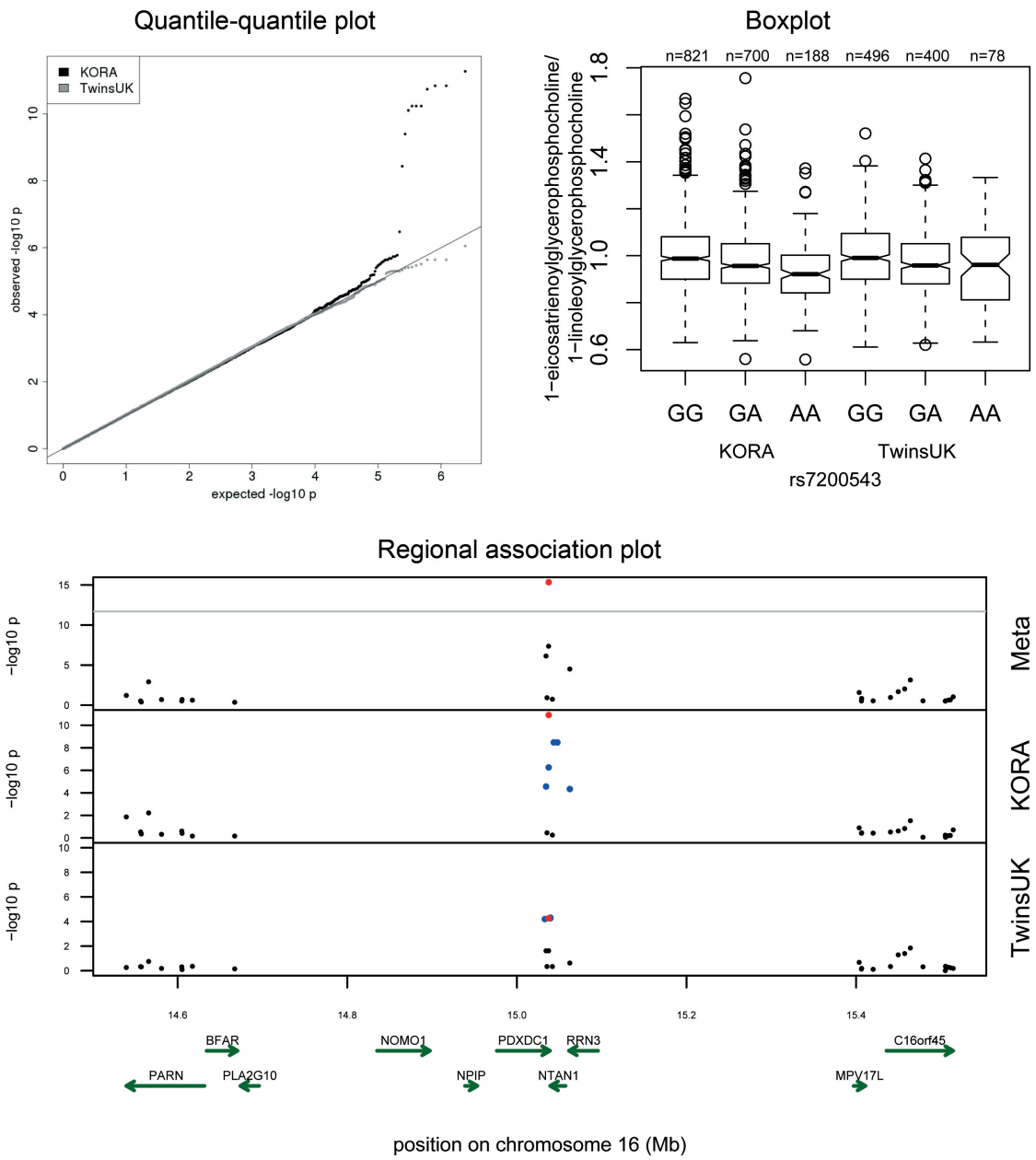

Figure A.1 (cont.)

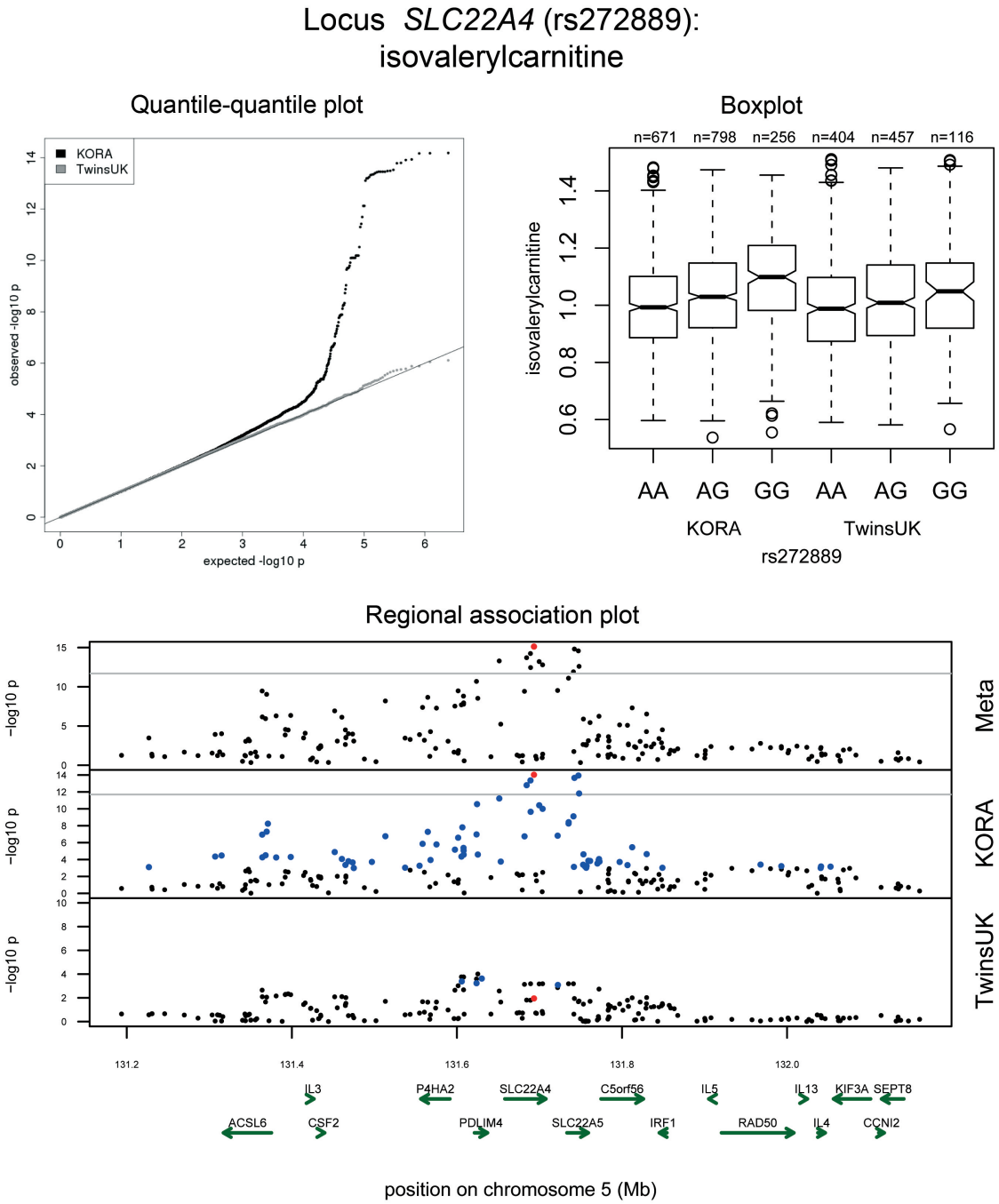

Figure A.1 (cont.)

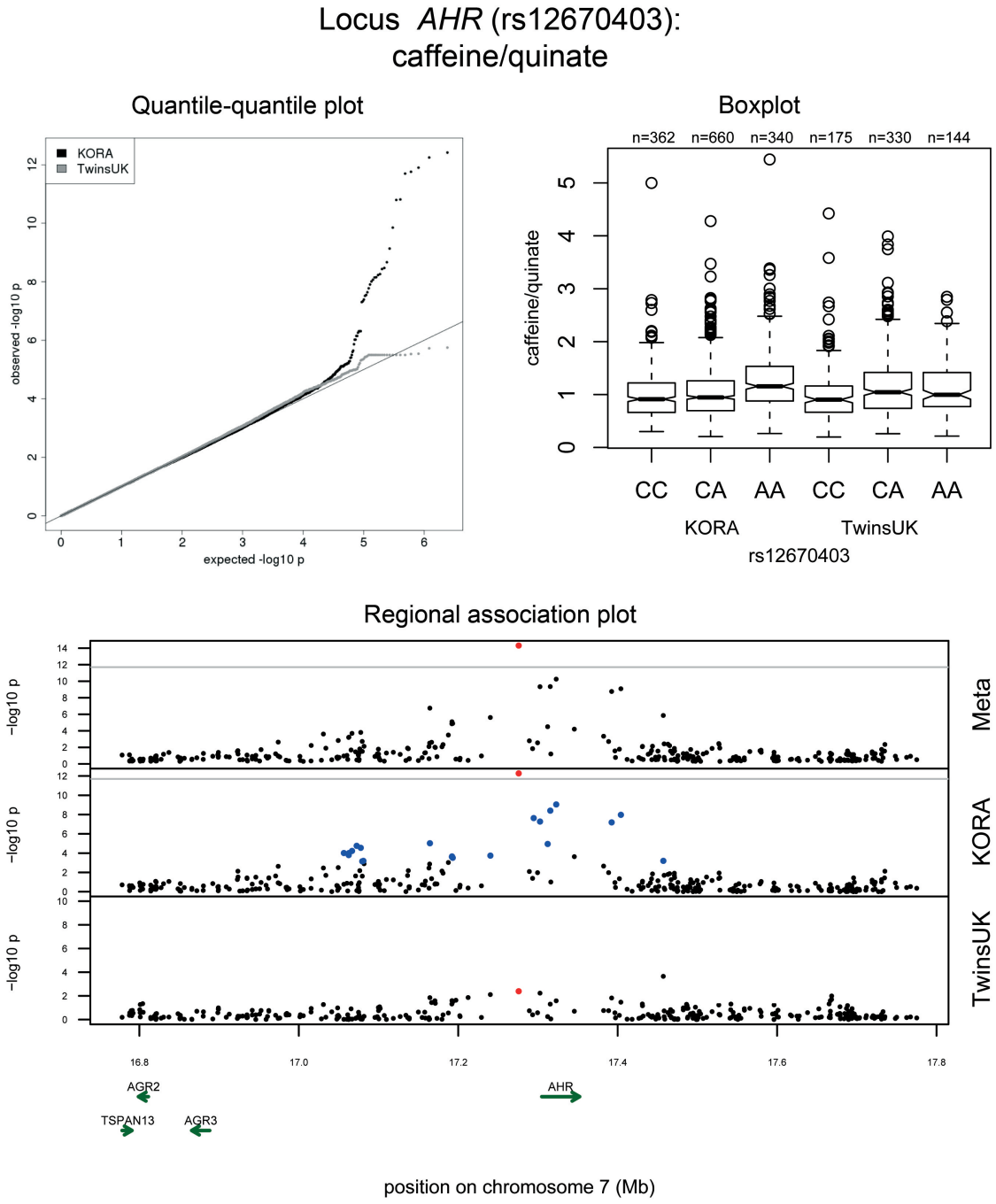

Figure A.1 (cont.)

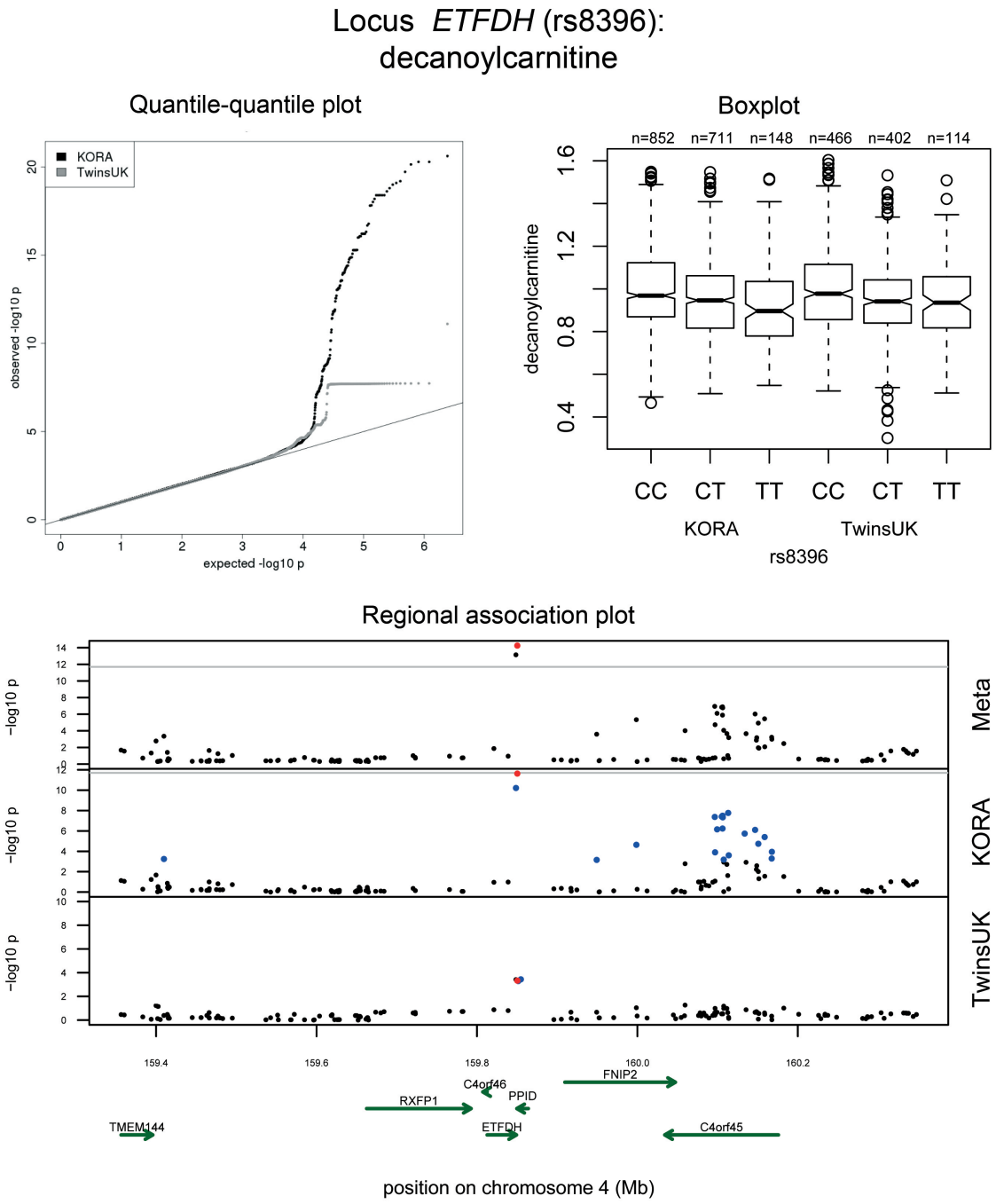

Figure A.1 (cont.)

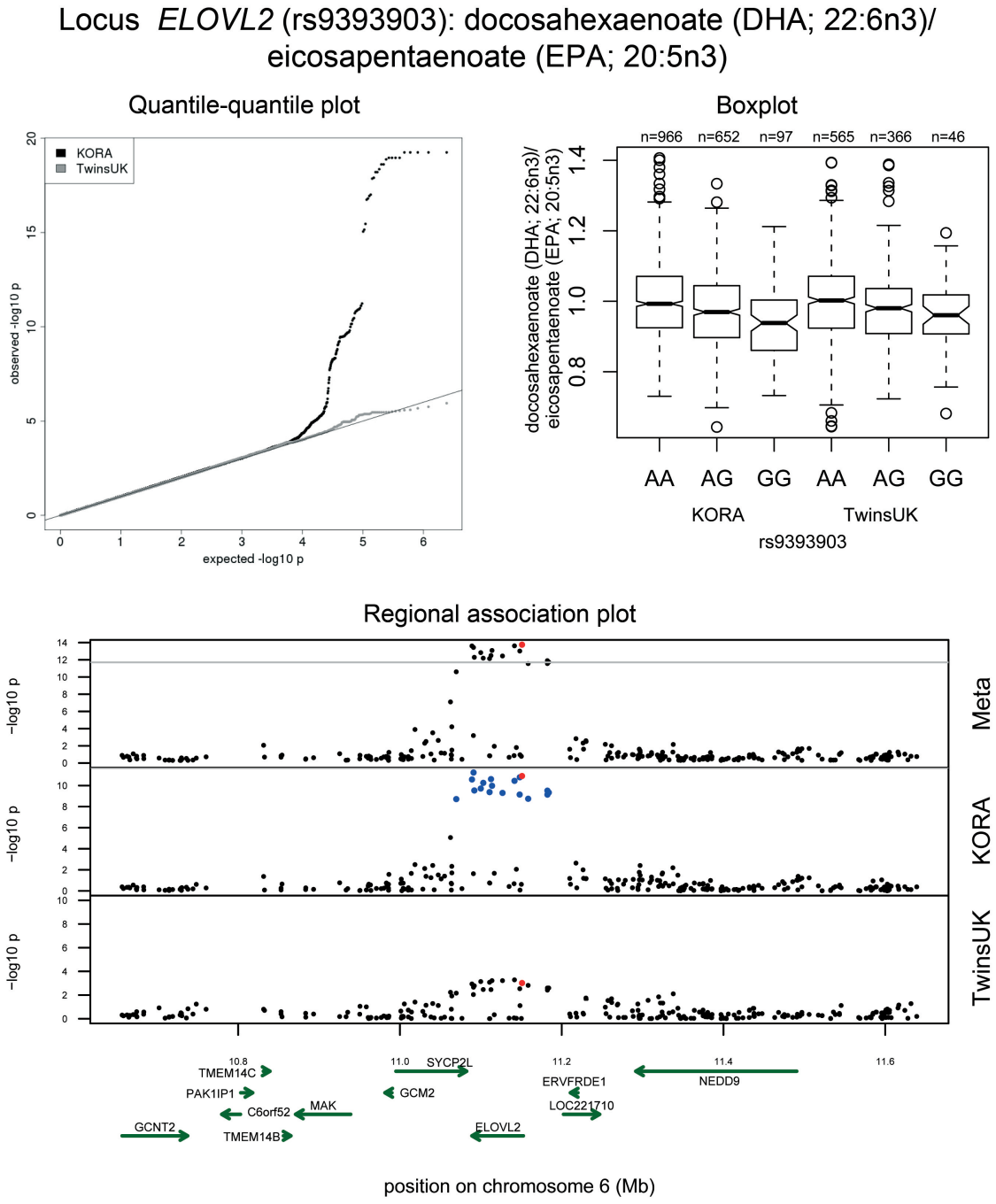

Figure A.1 (cont.)

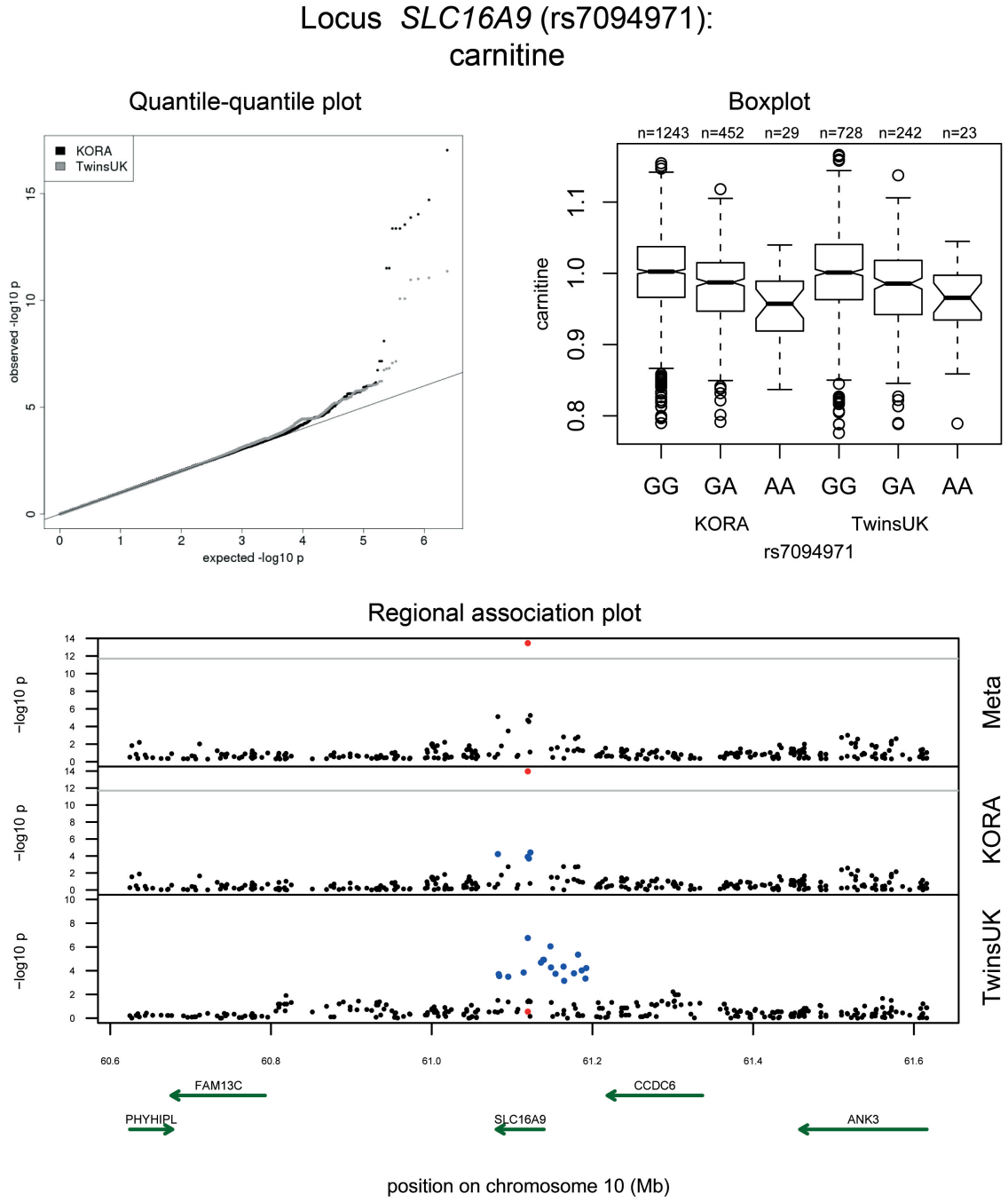

Figure A.1 (cont.)

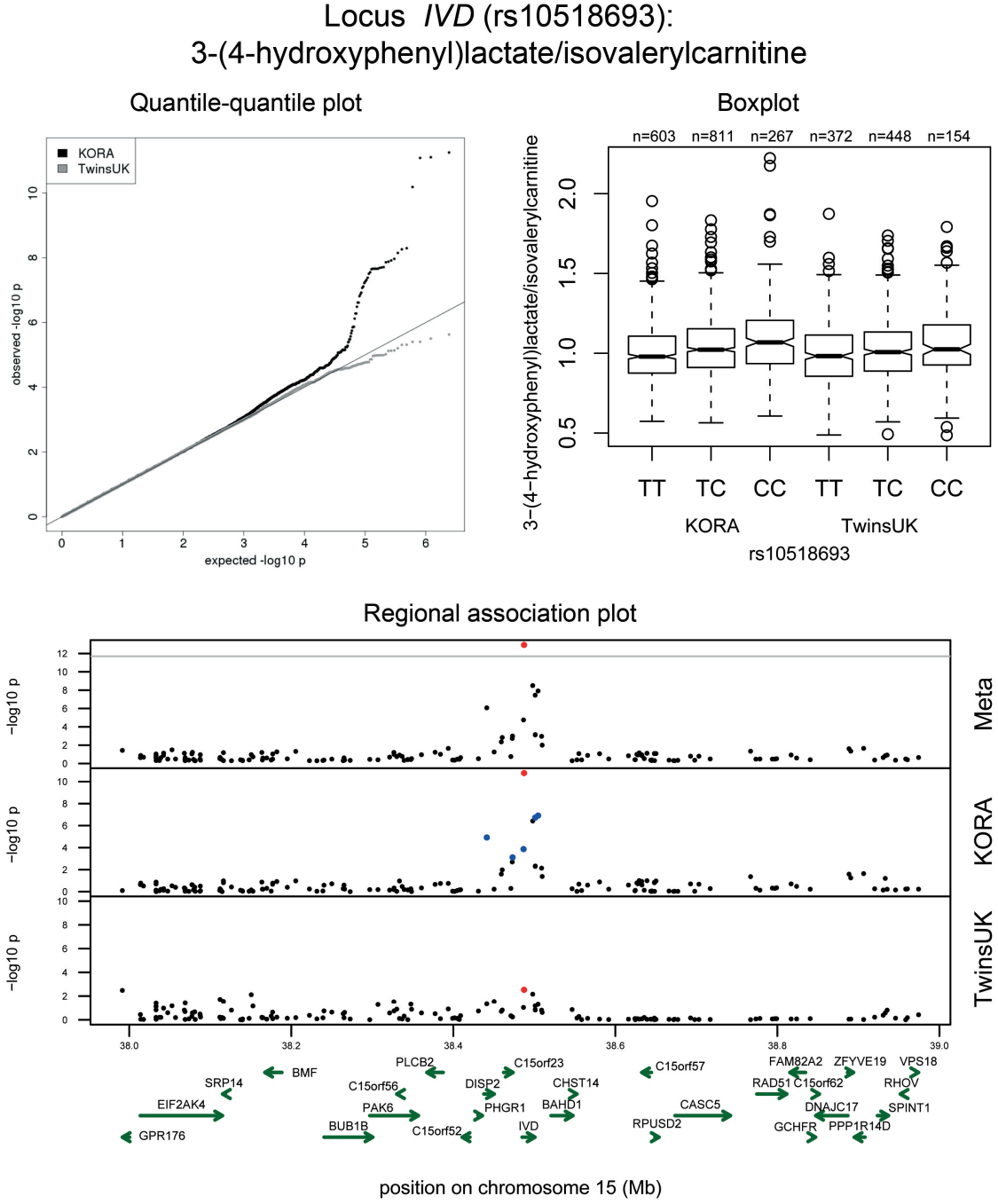

Figure A.1 (cont.)

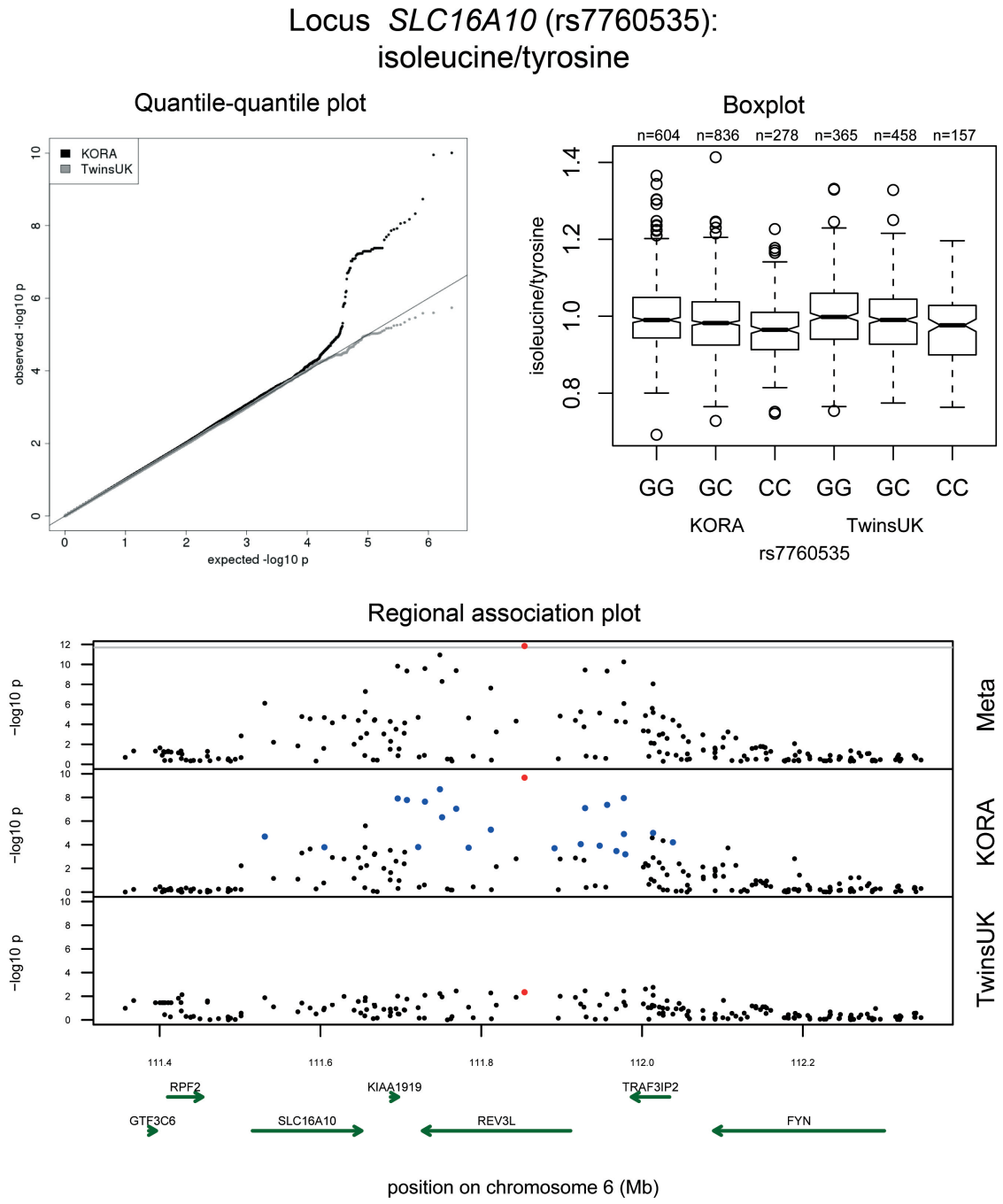

Figure A.1 (cont.)

References

Abecasis GR, Cherny SS, Cookson WO and Cardon LR, 2002. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101.

Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, Bueno-de-Mesquita HB, Gross M, Helzlsouer K, Jacobs EJ *et al.*, 2009. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet*, 41(9):986–990.

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

Asztalos BF, Cupples LA, Demissie S, Horvath KV, Cox CE, Batista MC and Schaefer EJ, 2004. High-density lipoprotein subpopulation profile and coronary heart disease prevalence in male participants of the Framingham Offspring Study. *Arterioscler Thromb Vasc Biol*, 24(11):2181–2187.

Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, Penninx BW, Janssens AC, Wilson JF, Spector T *et al.*, 2009. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet*, 41(1):47–55.

Barber MJ, Mangravite LM, Hyde CL, Chasman DI, Smith JD, McCarty CA, Li X, Wilke RA, Rieder MJ, Williams PT *et al.*, 2010. Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS One*, 5(3):e9763.

Barderas MG, Laborde CM, Posada M, de la Cuesta F, Zubiri I, Vivanco F and Alvarez-Llamas G, 2011. Metabolomic profiling for identification of novel potential biomarkers in cardiovascular diseases. *J Biomed Biotechnol*, 2011:790132.

Bielinski SJ, Chai HS, Pathak J, Talwalkar JA, Limburg PJ, Gullerud RE, Sicotte H, Klee EW, Ross JL, Kocher JP *et al.*, 2011. Mayo Genome Consortia: a

genotype-phenotype resource for genome-wide association studies with an application to the analysis of circulating bilirubin levels. *Mayo Clin Proc*, 86(7):606–614.

Boes E, Coassin S, Kollerits B, Heid IM and Kronenberg F, 2009. Genetic-epidemiological evidence on genes associated with HDL cholesterol levels: a systematic in-depth review. *Exp Gerontol*, 44(3):136–60.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ *et al.*, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Camont L, Chapman MJ and Kontush A, 2011. Biological activities of HDL subpopulations and their relevance to cardiovascular disease. *Trends Mol Med*, 17(10):594–603.

Castelli W, Doyle J, Gordon T, Hames C, Hjortland M, Hulley S, Kagan A and Zukel W, 1977. HDL cholesterol and other lipids in coronary heart disease. The cooperative lipoprotein phenotyping study. *Circulation*, 55(5):767–772.

Castelli WP, 1996. Lipids, risk factors and ischaemic heart disease. *Atherosclerosis*, 124(Suppl):S1–S9.

Chambers JC, Zhang W, Lord GM, van der Harst P, Lawlor DA, Sehmi JS, Gale DP, Wass MN, Ahmadi KR, Bakker SJ *et al.*, 2010. Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet*, 42(5):373–375.

Chambers JC, Zhang W, Sehmi J, Li X, Wass MN, van der Harst P, Holm H, Sanna S, Kavousi M, Baumeister SE *et al.*, 2011. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet*, 43(11):1131–1138.

Chasman DI, Paré G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, Hamsten A, Kathiresan S, Mälarstig A *et al.*, 2009. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet*, 5(11):e1000730.

Chen G, Ramos E, Adeyemo A, Shriner D, Zhou J, Doumatey AP, Huang H, Erdos MR, Gerry NP, Herbert A *et al.*, 2012. UGT1A1 is a major locus influencing bilirubin levels in African Americans. *Eur J Hum Genet*, 20(4):463–468.

Chung CM, Wang RY, Chen JW, Fann CS, Leu HB, Ho HY, Ting CT, Lin TH, Sheu SH, Tsai WC *et al.*, 2010. A genome-wide association study identifies new loci for ACE activity: potential implications for response to ACE inhibitor. *Pharmacogenomics J*, 10(6):abstract.

Corydon MJ, Andresen BS, Bross P, Kjeldsen M, Andreasen PH, Eiberg H, Kølvraa S and Gregersen N, 1997. Structural organization of the human short-chain acyl-CoA dehydrogenase gene. *Mamm Genome*, 8(12):922–926.

Danik JS, Paré G, Chasman DI, Zee RY, Kwiatkowski DJ, Parker A, Miletich JP and Ridker PM, 2009. Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study. *Circ Cardiovasc Genet*, 2(2):134–141.

De Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S and Voight BF, 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*, 17(R2):R122–R128.

Dehghan A, Dupuis J, Barbalic M, Bis JC, Eiriksdottir G, Lu C, Pellikka N, Wallaschofski H, Kettunen J, Henneman P *et al.*, 2011. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation*, 123(7):731–738.

Dehghan A, Yang Q, Peters A, Basu S, Bis JC, Rudnicka AR, Kavousi M, Chen MH, Baumert J, Lowe GD *et al.*, 2009. Association of novel genetic loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. *Circ Cardiovasc Genet*, 2(2):125–133.

Demirkan A, van Duijn C, Ugocsai P, Isaacs A, Pramstaller P, Liebisch G, Wilson J, Johansson A, Rudan I, Aulchenko Y *et al.*, 2012. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet*, 8(2):e1002490.

Devlin B and Roeder K, 1999. Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Döring A, Gieger C, Mehta D, Gohlke H, Prokisch H, Coassin S, Fischer G, Henke K, Klopp N, Kronenberg F *et al.*, 2008. SLC2A9 influences uric acid concentrations with pronounced sexspecific effects. *Nat Genet*, 40(4):430–436.

Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL *et al.*, 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42(2):105–116.

Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Smith AV, Tobin MD, Verwoert GC, Hwang SJ *et al.*, 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109.

Eijgelsheim M, Newton-Cheh C, Sotoodehnia N, de Bakker PI, Müller M, Morrison AC, Smith AV, Isaacs A, Sanna S, Dörr M *et al.*, 2010. Genome-wide association analysis identifies multiple loci related to resting heart rate. *Hum Mol Genet*, 19(19):3885–3894.

Evans AM, DeHaven CD, 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.

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R *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.

Frolkis A, Knox C, Lim E, Jewison T, Law V, Hau DD, Liu P, Gautam B, Ly S, Guo AC *et al.*, 2010. SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res*, 38(Database issue):D480–D487.

Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, Cohen W, Oudot-Mellakh T, Antoni G, Alessi MC *et al.*, 2011. Genetics of venous thrombosis: insights from a new genome wide association study. *PLoS One*, 6(9):e25581.

Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J *et al.*, 2008. Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, 4(11):e1000282.

Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, Serbanovic-Canic J, Elling U, Goodall AH, Labrune Y *et al.*, 2011. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208.

Hazra A, Kraft P, Selhub J, Giovannucci EL, Thomas G, Hoover RN, Chanock SJ and Hunter DJ, 2008. Common variants of FUT2 are associated with plasma vitamin B12 levels. *Nat Genet*, 40(10):1160–1162.

Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, Ugocsai P, Aulchenko Y, Franklin CS, Liebisch G, Erdmann J *et al.*, 2009. Genetic determinants of circulating sphingolipid concentrations in european populations. *PLoS Genet*, 5(10):e1000672.

Hindorff LA, MacArthur J, Wise A, Junkins HA, Hall PN, Klemm AK and Manolio TA, 2011. A catalog of published genome-wide association studies. `www.genome.gov/gwastudies`. Last accessed 16 February 2012.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS and Manolio TA, 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*, 106(23):9362–9367.

Holmes E, Wilson ID and Nicholson JK, 2008. Metabolic phenotyping in health and disease. *Cell*, 134(5):714–717.

Horgan RP, Clancy OH, Myers JE and Bakera PN, 2008. An overview of proteomic and metabolomic technologies and their application to pregnancy research. *BJOG*, 116(2):173–181.

Howie BN, Donnelly P and Marchini J, 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529.

Hsia DY, 1958. Phenylketonuria: the phenylalanine-tyrosine ratio in the detection of the heterozygous carrier. *J Ment Defic Res*, 2(1):8–16.

Huber F, Kalbitzer HR and Kremer W, 2005. Verfahren zur Bestimmung von Lipoproteinen in Körperflüssigkeiten und Messanordnung dafür. *DE 10 2004 026 903 B4*, Germany.

Huber F, Kalbitzer HR and Kremer W, 2011a. Process for the determination of lipoproteins in body fluids. *US7,927,878*, United Stated of America.

Huber F, Kalbitzer HR and Kremer W, 2011b. Process for the determination of lipoproteins in body fluids. *AU2005250571*, Australia.

Huuskonen J, Olkkonen VM, Jauhiainen M and Ehnholm C, 2001. The impact of phospholipid transfer protein (PLTP) on HDL metabolism. *Atherosclerosis*, 155(2):269–281.

Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmüller G, Kato BS, Mewes HW *et al.*, 2010. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*, 42(2):137–141.

Inouye M, Kettunen J, Soininen P, Silander K, Ripatti S, Kumpula LS, Hämäläinen E, Jousilahti P, Kangas AJ, Männistö S *et al.*, 2010. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol*, 6:441.

Jethva R, Bennett MJ and Vockley J, 2008. Short-chain acyl-coenzyme A dehydrogenase deficiency. *Mol Genet Metab*, 95(4):195–200.

Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, Martins RA, Kennedy BA, Hassell RG, Visser ME *et al.*, 2010. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet*, 42(8):684–687.

Johnson AD, Kavousi M, Smith AV, Chen MH, Dehghan A, Aspelund T, Lin JP, van Duijn CM, Harris TB, Cupples LA *et al.*, 2009. Genome-wide association meta-analysis for total serum bilirubin levels. *Hum Mol Genet*, 18(14):2700–2710.

Johnson T and Kutalik Z, 2008. QUICKTEST user guide. `http://toby.freeshell.org/software/quicktest-guide.pdf`. Last accessed 15 September 2011.

Kaess BM, Tomaszewski M, Braund PS, Stark K, Rafelt S, Fischer M, Hardwick R, Nelson CP, Debiec R, Huber F *et al.*, 2011. Large-scale candidate gene analysis of HDL particle features. *PLoS One*, 6(1):e14529.

Kane JP, Hardman DA and Paulus HE, 1980. Heterogeneity of apolipoprotein B: isolation of a new species from human chylomicrons. *Proc Natl Acad Sci*, 77(5):2465–2469.

Kanehisa M and Goto S, 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

Kanehisa M, Goto S, Furumichi M, Tanabe M and Hirakawa M, 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):D355–D360.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M and Hirakawa M, 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):D354–D357.

Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T *et al.*, 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet*, 41(1):56–65.

Kato N, Takeuchi F, Tabara Y, Kelly TN, Go MJ, Sim X, Tay WT, Chen CH, Zhang Y, Yamamoto K *et al.*, 2011. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat Genet*, 43(6):531–538.

Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikäinen LP, Kangas AJ, Soininen P, Würtz P, Silander K *et al.*, 2012. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet*, 44(3):269–276.

Kjolby M, Andersen OM, Breiderhoff T, Fjorback AW, Pedersen KM, Madsen P, Jansen P, Heeren J, Willnow TE and Nykjaer A, 2010. Sort1, encoded by the cardiovascular risk locus 1p13.3, is a regulator of hepatic lipoprotein export. *Cell Metab*, 12(3):213–223.

Koal T and Deigner HP, 2010. Challenges in mass spectrometry based targeted metabolomics. *Curr Mol Med*, 10(2):216–226.

Kojadinovic I and Yan J, 2010. Modeling multivariate distributions with continuous margins using the copula R package. *J Stat Softw*, 34(9):1–20.

Kolz M, Johnson T, Sanna S, Teumer A, Vitart V, Perola M, Mangino M, Albrecht E, Wallace C, Farrall M *et al.*, 2009. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet*, 5(6):e1000504.

Kraja AT, Vaidya D, Pankow JS, Goodarzi MO, Assimes TL, Kullo IJ, Sovio U, Mathias RA, Sun YV, Franceschini N *et al.*, 2011. A bivariate genome-wide ap-

proach to metabolic syndrome: STAMPEED consortium. *Diabetes*, 60(4):1329–1339.

Krug S, Kastenmüller G, Stückler F, Rist MJ, Skurk T, Sailer M, Raffler J, Römisch-Margl W, Adamski J, Prehn C *et al.*, 2012. The dynamic range of the human metabolome revealed by challenges. *FASEB J*, page [Epub ahead of print].

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

Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV *et al.*, 2010. New loci associated with kidney function and chronic kidney disease. *Nat Genet*, 42(5):376–384.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S *et al.*, 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838.

Lattka E, Illig T, Koletzko B and Heinrich J, 2010. Genetic variants of the FADS1 FADS2 gene cluster as related to essential fatty acid metabolism. *Curr Opin Lipidol*, 21(1):64–69.

Lemaitre RN, Tanaka T, Tang W, Manichaikul A, Foy M, Kabagambe EK, Nettleton JA, King IB, Weng LC, Bhattacharya S *et al.*, 2011. Genetic loci associated with plasma phospholipid n-3 fatty acids: A meta-analysis of genome-wide association studies from the CHARGE consortium. *PLoS Genet*, 7(7):e1002193.

Levey AS, Coresh J, Greene T, Marsh J, Stevens LA, Kusek JW and Van Lente F, 2007. Expressing the modification of diet in renal disease study equation for estimating glomerular filtration rate with standardized serum creatinine values. *Clin Chem*, 53(4):766–772.

Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T *et al.*, 2009. Genome-wide association study of blood pressure and hypertension. *Nat Genet*, 41(6):677–687.

Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon *et al.*, 2007. Novel Crohn disease locus identified

by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet*, 3(4):e58.

Linsel-Nitschke P, Jansen H, Aherrarhou Z, Belz S, Mayer B, Lieb W, Huber F, Kremer W, Kalbitzer HR, Erdmann J *et al.*, 2009. Macrophage cholesterol efflux correlates with lipoprotein subclass distribution and risk of obstructive coronary artery disease in patients undergoing coronary angiography. *Lipids Health Dis*, 8:14.

Lovely RS, Yang Q, Massaro JM, Wang J, D'Agostino RB, O'Donnell CJ, Shannon J and Farrell DH, 2011. Assessment of genetic determinants of the association of γ fibrinogen in relation to cardiovascular disease. *Arterioscler Thromb Vasc Biol*, 31(10):2345–2352.

Lu X, Zhao W, Huang J, Li H, Yang W, Wang L, Huang W, Chen S and Gu D, 2007. Common variation in KLKB1 and essential hypertension risk: tagging-SNP haplotype analysis in a case-control study. *Hum Genet*, 121(3-4):327–335.

Maher B, 2008. The case of the missing heritability. *Nature*, 456(6):18–21.

Maier EM, Liebl B, Roschinger W, Nennstiel-Ratzel U, Fingerhut R, Olgemoller B, Busch U, Krone N, v Kries R and Roscher AA, 2005. Population spectrum of ACADM genotypes correlated to biochemical phenotypes in newborn screening for medium-chain acyl-CoA dehydrogenase deficiency. *Hum Mutat*, 25(5):443–452.

Malaisse WJ, Malaisse-Lagae F, Davies DR, Vandercammen A and van Schaftingen E, 1990. Regulation of glucokinase by a fructose-1-phosphate-sensitive protein in pancreatic islets. *Eur J Biochem*, 190(3):539–545.

Malet-Martino M and Holzgrabe U, 2011. NMR techniques in biomedical and pharmaceutical analysis. *J Pharm Biomed Anal*, 55(1):1–15.

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

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP and Hirschhorn JN, 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369.

McGovern DP, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M, Ippoliti A, Vasiliauskas E, Berel D, Derkowski C *et al.*, 2010. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum Mol Genet*, 19(17):3468–3476.

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM *et al.*, 2010. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719.

Nagrath D, Caneba C, Karedath T and Bellance N, 2011. Metabolomics for mitochondrial and cancer studies. *Biochim Biophys Acta*, 1807(6):650–663.

Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S *et al.*, 2009. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet*, 41(6):666–676.

Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW *et al.*, 2011. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet*, 7(9):e1002270.

Nicholson JK and Lindon JC, 2008. Systems biology: Metabonomics. *Nature*, 455(7216):1054–1056.

Okada Y, Kamatani Y, Takahashi A, Matsuda K, Hosono N, Ohmiya H, Daigo Y, Yamamoto K, Kubo M, Nakamura Y *et al.*, 2010. A genome-wide association study in 19 633 Japanese subjects identified LHX3-QSOX2 and IGF1 as adult height loci. *Hum Mol Genet*, 19(11):2303–2312.

Panneerselvam K and Freeze HH, 1996. Mannose enters mammalian cells using a specific transporter that is insensitive to glucose. *J Biol Chem*, 271(16):9417–9421.

Paradis E, Claude J and Strimmer K, 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

Pearson TA and Manolio TA, 2008. How to interpret a genome-wide association study. *JAMA*, 299(11):1335–1344.

Petersen AK, Stark K, Musameh MD, Nelson CP, Römisch-Margl W, Kremer W, Raffler J, Krug S, Skurk T, Rist MJ *et al.*, 2012. Genetic associations with

lipoprotein subfractions provide information on their biological nature. *Hum Mol Genet*, 21(6):1433–1443.

Plomin R, Haworth CM and Davis OS, 2009. Common disorders are quantitative traits. *Nat Rev Genet*, 10(12):872–878.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al.*, 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575.

R Development Core Team, 2010. R: A language and environment for statistical computing. Vienna, Austria.

Rader DJ, 2006. Molecular regulation of HDL metabolism and function: implications for novel therapies. *J Clin Invest*, 116(12):3090–3100.

Rader DJ, 2009. Lecithin: cholesterol acyltransferase and atherosclerosis: another high-density lipoprotein story that doesn't quite follow the script. *Circulation*, 120(7):549–552.

Reilly MP, Li M, He J, Ferguson JF, Stylianou IM, Mehta NN, Burnett MS, Devaney JM, Knouff CW, Thompson JR *et al.*, 2011. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *Lancet*, 377(9763):383–392.

Richards JB, Rivadeneira F, Inouye M, Pastinen TM, Soranzo N, Wilson SG, Andrew T, Falchi M, Gwilliam R, Ahmadi KR *et al.*, 2008. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet*, 371(9623):1505–1512.

Ridker PM, Pare G, Parker A, Zee RY, Danik JS, Buring JE, Kwiatkowski D, Cook NR, Miletich JP and Chasman DI, 2008. Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am J Hum Genet*, 82(5):1185–1192.

Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, van den Berg D, Matullo G, Baris D *et al.*, 2010. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet*, 42(11):978–984.

Sabatti C, Service SK, Hartikainen AL, Poutam A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M *et al.*, 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, 41(1):35–46.

Samani NJ, Braund PS, Erdmann J, Gotz A, Tomaszewski M, Linsel-Nitschke P, Hajat C, Mangino M, Hengstenberg C, Stark K *et al.*, 2008. The novel genetic variant predisposing to coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol. *J Mol Med*, 86(11):1233–1241.

Sanna S, Busonero F, Maschio A, McArdle PF, Usala G, Dei M, Lai S, Mulas A, Piras MG, Perseu L *et al.*, 2009. Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum Mol Genet*, 18(14):2711–2718.

Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, Jackson AU *et al.*, 2010. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet*, 42(2):142–148.

Shimodaira H, 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*, 51(3):492–508.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S *et al.*, 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885.

Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, Li M *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 EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Mägi R *et al.*, 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11):937–948.

Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wägele B, Altmaier E, CARDIoGRAM, Deloukas P, Erdmann J *et al.*, 2011a. Human metabolic

individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60.

Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D *et al.*, 2011b. A genome-wide association study of metabolic traits in human urine. *Nat Genet*, 43(6):565–569.

Suzuki R and Shimodaira H, 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542.

Taguchi T, Yamashita E, Mizutani T, Nakajima H, Yabuuchi M, Asano N and Miwa I, 2005. Hepatic glycogen breakdown is implicated in the maintenance of plasma mannose concentration. *Am J Physiol Endocrinol Metab*, 288(3):E534–E540.

Tanaka T, Scheet P, Giusti B, Bandinelli S, Piras MG, Usala G, Lai S, Mulas A, Corsi AM, Vestrini A *et al.*, 2009a. Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations. *Am J Hum Genet*, 84(4):477–482.

Tanaka T, Shen J, Abecasis GR, Kisialiou A, Ordovas JM, Guralnik JM, Singleton A, Bandinelli S, Cherubini A, Arnett D *et al.*, 2009b. Genome-wide association sudy of plasma polyunsaturated fatty acids in the InCHIANTI study. *PLoS Genet*, 5(1):e1000338.

Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP and Clark TG, 2007. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, 23(20):2741–274.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ *et al.*, 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713.

Teupser D, Baber R, Ceglarek U, Scholz M, Illig T, Gieger C, Holdt LM, Leichtle A, Greiser KH, Huster D *et al.*, 2010. Genetic regulation of serum phytosterol levels and risk of coronary artery disease. *Circ Cardiovasc Genet*, 3(4):331–339.

Thompson JR, Attia J and Minelli C, 2011. The meta-analysis of genome-wide association studies. *Brief Bioinform*, 12(3):259–269.

Tomaszewski M, Debiec R, Braund PS, Nelson CP, Hardwick R, Christofidou P, Denniff M, Codd V, Rafelt S, van der Harst P *et al.*, 2010. Genetic architecture

of ambulatory blood pressure in the general population: insights from cardiovascular genecentric array. *Hypertension*, 56(6):1069–1076.

Trégouët DA, Heath S, Saut N, Biron-Andreani C, Schved JF, Pernod G, Galan P, Drouet L, Zelenika D, Juhan-Vague I *et al.*, 2009. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood*, 113(21):5298–5303.

Tukiainen T, Kettunen J, Kangas AJ, Lyytikäinen LP, Soininen P, Sarin AP, Tikkanen E, O'Reilly MJ P Fand Savolainen, Kaski K, Pouta A *et al.*, 2012. Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Hum Mol Genet*, 21(6):1444–1455.

Van Schaftingen E, 1989. A protein from rat liver confers to glucokinase the property of being antagonistically regulated by fructose 6-phosphate and fructose 1-phosphate. *Eur J Biochem*, 179(1):179–184.

Von Eckardstein A, 2010. Implications of torcetrapib failure for the future of HDL therapy: is HDL-cholesterol the right target? *Expert Rev Cardiovasc Ther*, 8(3):345–358.

Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, Ahmadi K, Dobson RJ, Marçano AC, Hajat C *et al.*, 2008. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet*, 82(1):139–149.

Warner JP, Leek JP, Intody S, Markham AF and Bonthron DT, 1995. Human glucokinase regulatory protein (GCKR): cDNA and genomic cloning, complete primary structure, and chromosomal localization. *Mamm Genome*, 6(8):532–536.

Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, Ripatti S, Aulchenko YS, Zhang W, Yuan X, Lim N *et al.*, 2010. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol*, 30(11):2264–2276.

Whitfield JB and Martin NG, 1983. Determinants of variation in plasma alkaline phosphatase activity: A twin study. *Am J Hum Genet*, 35(5):978–986.

Wichmann HE, Gieger C and Illig T, 2005. KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 67(Suppl 1):S26–S30.

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM *et al.*, 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*, 40(2):161–169.

Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S *et al.*, 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res*, 37(Database issue):D603–610.

World Health Organisation, 2011a. Cardiovascular diseases (CVDs). Facts sheet No. 317. `http://www.who.int/mediacentre/factsheets/fs317/en/index.html`. Last accessed November 11th 2011.

World Health Organisation, 2011b. Diabetes. Fact sheet No. 312. `http://www.who.int/mediacentre/factsheets/fs312/en/index.html`. Last accessed November 11th 2011.

Yan J, 2007. Enjoy the joy of copulas: With a package copula. *J Stat Softw*, 21(4):1–21.

Yanagisawa Y, Nishimura H, Matsuki H, Osaka F and Kasuga H, 1986. Personal exposure and health effect relationship for NO2 with urinary hydroxyproline to creatinine ratio as indicator. *Arch Environ Health*, 41(1):41–48.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al.*, 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–569.

Yazdanyar A, Yeang C and Jiang XC, 2011. Role of phospholipid transfer protein in high-density lipoprotein- mediated reverse cholesterol transport. *Curr Atheroscler Rep*, 13(3):242–248.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N *et al.*, 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*, 39(5):645–649.

Yuan X, Waterworth D, Perry JR, Lim N, Song K, Chambers JC, Zhang W, Vollenweider P, Stirnadel H, Johnson T *et al.*, 2008. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet*, 83(4):520–528.

Zabaneh D and Balding DJ, 2010. A genome-wide association study of the metabolic syndrome in Indian Asian men. *PLoS One*, 5(8):e11961.

Zeggini E and Ioannidis J, 2009. Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10(2):191–201.

Zhai G, Teumer A, Stolk L, Perry JR, Vandenput L, Coviello AD, Koster A, Bell JT, Bhasin S, Eriksson J *et al.*, 2011. Eight common genetic variants associated with serum DHEAS levels suggest a key role in ageing mechanisms. *PLoS Genet*, 7(4):e1002025.

Acknowledgements

First, I would like to thank my supervisor Prof. Dr. Dr. H.-Erich Wichmann who gave me the possibility to realise this work. Special thanks go to my co-supervisor Dr. Christian Gieger who provided the idea of this thesis and who always found time to support my work and to give helpful advice despite his tight schedule. I also thank Prof. Dr. Karsten Suhre for fruitful discussions about the evaluation of metabolomics data and Dr. Nicole Soranzo for giving me the opportunity to do an exchange with the Wellcome Trust Sanger Institute. Moreover, I thank all my colleagues of the Institute of Genetic Epidemiology at the Helmholtz Zentrum München for their professional advice. I would also like to thank Carolin, Sigrid, Marcus, Anja, Michael and Henrike to whom I owe many good memories of this time.

Nevertheless, my biggest thanks go to my family who supported me in all situations.