Year : 2010

Computational analysis of the genetic and environmental contributions to disease-related human phenotypes:
From predicting adverse lipid response of HIV patients receiving antiretroviral therapy to genome-wide association studies of cardiovascular and lipid related disorders

Diana MAREK

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Département de Génétique Médicale - Computational Biology Group

Computational analysis of the genetic and environmental contributions to disease-related human phenotypes:

From predicting adverse lipid response of HIV patients receiving antiretroviral therapy to genome-wide association studies of cardiovascular and lipid-related disorders

Thèse de doctorat ès sciences de la vie (PhD)

Présentée à la Faculté de biologie et médecine de l'Université de Lausanne par

DIANA MAREK

Diplômée de l'Université Paris XI – Orsay (France)
Master de Bioinformatique et Biostatistiques

Jury

Prof. Jérôme Biollaz, Président
Prof. Sven Bergmann, Directeur de thèse
Prof. Jacques Beckmann, co-Directeur de thèse
Prof. Murielle Bochud, Expert
Prof. Eckart Zitzler, Expert

Lausanne, 2010

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur	Prof.	Jérôme **Biollaz**
Directeur de thèse	Monsieur	Prof.	Sven **Bergmann**
Co-directeur de thèse	Monsieur	Prof.	Jacques **Beckmann**
Experts	Madame	Dr	Murielle **Bochud**
	Monsieur	Prof.	Echart **Zitzler**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Diana Marek

Master en Bioinformatique de l'Université Orsay-Paris XI, France

intitulée

Analyse bioinformatique et statistique des contributions génétiques et environnementales liées aux désordres lipidiques chez l'humain.

Lausanne, le 22 octobre 2010

pour Le Doyen
de la Faculté de Biologie et de Médecine

Prof. Jérôme Biollaz

Jérôme Goudet

Ici et ailleurs,
A mes parents,
A tous ceux et celles qui ont marqué ma vie,
De quelque manière qui soit.

Acknowledgements

I spent five years in the Department of Medical Genetics at the University of Lausanne for my PhD. Since 2005, I met many people at work and in my life in Lausanne. They all participated in different ways to a very special step in my scientific career and life as a human being.

I would like to start with the Department of Medical Genetics (DGM). I saw the DGM growing. I appreciated being one of the first DGM member and the first student of the Computational Biology Group (CBG) founded by Professor Sven Bergmann.

I would like to thank the Swiss Institute of Bioinformatics (SIB). The CBG is affiliated to the SIB. It allowed me to meet other scientists from the community, to attend conferences, and especially through the SIB Doctoral School program that I joined to meet other PhD students, and share my experience with them. It has been a fantastic way of knowing what other bioinformaticians are doing in Switzerland and abroad. Thanks to all of you!

Through my PhD, I established collaborations with different research groups. I want here to thank them for giving me the opportunity to work with them.
First, thanks to Professor Amalio Telenti and Dr Philip Tarr who provided me with the Swiss HIV Cohort Study data and for the fruitful exchanges we had. Working on that topic was very stimulating for me.
Thanks to the Cohorte of Lausanne (CoLaus) Study, especially Vincent Mooser (from GSK), Peter Vollenweider and Gérard Waeber (from the CHUV) for setting up such a rich study that allowed me to be part of important large-scale projects which brought together lots of scientists, clinicians and researchers from all around the world. I learnt a lot and I developed my sense of collaboration and teamwork within two big consortia. Thanks to the different members of GIANT and MAGIC. It was very pleasant and rewarding working and exchanging with you.
Thanks to the Hypergenes project led by Dr. Carlo Rivolta. I have been involved in the primary steps of this study, which was a different and profitable experience for me.
I would also like to thank Dr. Ronan Roussel for the nice collaboration that we established together.

In the following section, I want to thank my experts: Prof. Eckart Zitzler and Prof. Murielle Bochud. Writing a thesis takes time but reading all those pages also demands time. Thanks to both of you for accepting to be my experts and for your involvement. Your suggestions, comments and questions have been very much appreciated.

During my PhD, my work in the CBG has been supervised by Prof. Sven Bergmann and Prof. Jacques Beckmann. I would like to thank Sven for his patience, advices, open-mindness and support in the good and bad moments of my PhD. I could easily share with you my fears and frustrations regarding my work. It was our first experience as PhD student and supervisor, so we certainly made mistakes but we learnt a lot from each other. I enjoyed working at the CBG with you and the rest of the group. I also want to say a few words about Jacqui. He was always available when needed for scientific and non-scientific matters. I always felt very comfortable talking to you; you pushed me over my limits because you believed I could do it. Both Sven and Jacqui boosted my self-confidence and challenged me to go forward when I wanted to give up. Thanks to the two of you for your encouragement. It helped me a lot all along the way!

I have been extremely lucky to be surrounded at work by very smart and nice people. I want to thank here all my colleagues at the CBG for their help and the time they spent with me for questions and discussions. Especially thanks to Zoltán, Bastian, Aitana, Micha, Armand, and Karen. Working with you guys was really great! Aitana, sharing our frustrations and good moments together at work made me realize you are a very special person to me.
I also want to thank the whole DGM for all the moments shared together, all the enthusiasm people showed in the department. I had lot of fun working in such a nice environment. I also thank Suzanne, our secretary. She has always been available for any administrative issue, I really appreciated your help.

I met lots of people during my PhD time and some of them became good friends. Here I am grateful to them and to my long-time friends. I tried my best not to forget anybody but I am sure that I will. So please, forgive me in advance!

Merci à mes amis de plus ou moins longues dates, à ceux que j'ai connus aux différentes époques de la vie, comme Marion et Fred à l'école primaire et en fac. Merci aux étudiants de la filière de Bionfo et Biostats de la fac d'Orsay. Ce furent 3 années géniales. Merci

Christophe, merci de m'avoir soutenue durant ces années, que ce soit pour les cours, les TPs, les examens, les coups durs de la vie. Tu as toujours été là, pour m'écouter, rigoler, bosser, et pour m'encourager !
Merci Pierrot pour tous les moments, conversations diverses et variées, partagés au long de nos thèses. Cela m'a permis de continuer le dur cheminement du doctorat.
Merci à toi Julien, avec qui j'ai repris contact ces dernières années, nos échanges et nos rencontres autour d'un café sont toujours géniales.
Merci à toi Jean, je t'ai connu vers le début de la fin de ma thèse, certainement au moment le plus dur. Tu m'as beaucoup apporté, ton franc-parler, nos convers', nos rigolades et ta façon bien propre à toi de me soutenir m'ont beaucoup aidée.
Merci au volley-ball, à ce sport qui est une passion pour moi depuis déjà 16 ans. Il m'a permis de me défouler lorsque j'en avais besoin. Il m'a permis de me vider la tête lorsque je pensais ne plus pouvoir y arriver. Merci à Manon avec qui je prends toujours autant de plaisir à jouer à ce sport.
Merci à Sandrine, Gérald, Célia et Sacha, mes Suisses "préférés". Partager le volley, faire les cours de soutien, de bonnes bouffes et des sorties avec vous m'ont fait découvrir une super famille. Je vous ai adoptés comme vous m'avez adoptée. Je suis ravie d'avoir croisé votre chemin et j'en garde d'excellents souvenirs.
Merci à Stelle, Marie-Aude, François, Pierre D, Julie, Séverine, Aurélie pour votre bonne humeur, votre écoute, vos encouragements à différents moments de ma thèse.
Gracias a vos Marcos, por haberme soportado todos estos años, y haber compartido conmigo momentos de mi vida más o menos fáciles y siempre haberme dado ánimo.
Thanks to Arne and David for your encouragement when I was deeply thinking whether I should start a PhD. Well, now it is done, and I will never forget your support in this special moment.
Thanks to Nick. Meeting you at the DGM was one of the best things that happened to me during those years. So many moments and slices of life shared here in Lausanne. My only regret is that I did not speak enough in French with you.

Merci à toi Valérie. Je sens déjà que les mots me manquent. Val, plus de 4 ans d'amitié. Je ne pourrai jamais oublier le début de celle-ci. Et depuis ça n'a pas cessé. On a fait Lausanne ensemble Louloute et ça c'est gravé à jamais. Les rires, les fous rires, les sourires, les pleurs, les coups durs, les moments de réconfort, les moments de stress, les voyages, les sorties, les dîners, les soirées, les organisations, les pauses café, les blagues, les quiproquos, les cartes…

c'est juste énorme ! Un livre louloute, un livre il faut écrire ! Merci, merci, merci. Tu m'as soutenue à chaque instant, j'ai toujours pu compter sur toi et cela n'a pas de prix ! Ma thèse sans toi n'aurait jamais été la même ! Je t'aime fort.

Merci à Jean-Jacques, dit Gigi. Cinq ans de thèse, cinq ans à te côtoyer ! Destin ? Coïncidence ? Le mystère est entier. Mais comme j'aime à dire, rien n'est dû au hasard et la vie ne cessera de m'émerveiller. Nos chemins devaient sûrement se croiser. On en a partagé des moments et surmonté des épreuves en cinq ans, pas toujours faciles et oui, car la vie n'est pas un long fleuve tranquille (c'est ça qu'on aime !) mais une chose est sûre, tu t'es toujours montré à l'écoute lorsque j'en ressentais le besoin et tu n'as jamais cessé de croire en moi, de me redonner courage et confiance quand je perdais espoir. Je t'aime.

Enfin, j'aimerais remercier ma famille, une famille multinationale et pluriculturelle. Sans nul doute responsable de mon ouverture d'esprit et de mon incessante envie de connaître, rencontrer, m'enrichir. Gracias a mi familia en Argentina. Mais surtout merci à mes parents. Vous m'avez chacun à votre manière inculqué les valeurs essentielles et avec lesquelles j'ai grandi. Maman, j'ai toujours pu compter sur toi, sur tes conseils souvent très justes, et surtout sur ton amour. Gracias por ser una madre tan especial y haber estado a mi lado hasta hoy. Papa, j'aurais tellement voulu partager ce moment avec toi, mais malgré la distance physique, jamais tu n'as été loin de mon cœur. Je vous aime tous les deux.

Cette thèse est à mon sens à la fois la fin et le début du chemin. La fin car depuis mes 15 ans, les évènements et ma vision utopiste de la vie m'ont conduite à vouloir entrer dans le monde de la recherche. C'est chose faite. Cette expérience m'a énormément enrichie tant sur le plan professionnel que personnel et je ne regrette à aucun instant mon choix même si le cheminement n'a pas été des plus simples. Ce doctorat représente également un commencement, un nouveau départ, un chemin où j'ai envie de mettre en valeur tout ce que j'ai pu apprendre en science et sur moi-même. Un chemin où mes aspirations et mes rêves pourront se développer. Et vous y avez participé ! Alors un grand merci à vous tous !

Abstract

Genetic variants influence the risk to develop certain diseases or give rise to differences in drug response. Recent progresses in cost-effective, high-throughput genome-wide techniques, such as microarrays measuring Single Nucleotide Polymorphisms (SNPs), have facilitated genotyping of large clinical and population cohorts. Combining the massive genotypic data with measurements of phenotypic traits allows for the determination of genetic differences that explain, at least in part, the phenotypic variations within a population. So far, models combining the most significant variants can only explain a small fraction of the variance, indicating the limitations of current models. In particular, researchers have only begun to address the possibility of interactions between genotypes and the environment. Elucidating the contributions of such interactions is a difficult task because of the large number of genetic as well as possible environmental factors.

In this thesis, I worked on several projects within this context. My first and main project was the identification of possible SNP-environment interactions, where the phenotypes were serum lipid levels of patients from the Swiss HIV Cohort Study (SHCS) treated with antiretroviral therapy. Here the genotypes consisted of a limited set of SNPs in candidate genes relevant for lipid transport and metabolism. The environmental variables were the specific combinations of drugs given to each patient over the treatment period. My work explored bioinformatic and statistical approaches to relate patients' lipid responses to these SNPs, drugs and, importantly, their interactions. The goal of this project was to improve our understanding and to explore the possibility of predicting dyslipidemia, a well-known adverse drug reaction of antiretroviral therapy. Specifically, I quantified how much of the variance in lipid profiles could be explained by the host genetic variants, the administered drugs and SNP-drug interactions and assessed the predictive power of these features on lipid responses. Using cross-validation stratified by patients, we could not validate our hypothesis that models that select a subset of SNP-drug interactions in a principled way have better predictive power than the control models using "random" subsets. Nevertheless, all models tested containing SNP and/or drug terms, exhibited significant predictive power (as compared to a random predictor) and explained a sizable proportion of variance, in the patient *stratified* cross-validation context. Importantly, the model containing stepwise selected SNP terms showed higher capacity to predict triglyceride levels than a model containing randomly selected

SNPs. Dyslipidemia is a complex trait for which many factors remain to be discovered, thus missing from the data, and possibly explaining the limitations of our analysis. In particular, the interactions of drugs with SNPs selected from the set of candidate genes likely have small effect sizes which we were unable to detect in a sample of the present size (<800 patients).

In the second part of my thesis, I performed genome-wide association studies within the Cohorte Lausannoise (CoLaus). I have been involved in several international projects to identify SNPs that are associated with various traits, such as serum calcium, body mass index, two-hour glucose levels, as well as metabolic syndrome and its components. These phenotypes are all related to major human health issues, such as cardiovascular disease. I applied statistical methods to detect new variants associated with these phenotypes, contributing to the identification of new genetic loci that may lead to new insights into the genetic basis of these traits. This kind of research will lead to a better understanding of the mechanisms underlying these pathologies, a better evaluation of disease risk, the identification of new therapeutic leads and may ultimately lead to the realization of "personalized" medicine.

Résumé

La recherche en génétique est en pleine expansion. L'analyse de grandes quantités de données génétiques, comme les "Single Nucleotide Polymorsphisms" (SNPs), combinées aux mesures phénotypiques au sein d'une population permet de détecter des marqueurs associés à des variations phénotypiques, telles que le risque de développer une maladie ou la différence de réponse à un traitement. Ces variants génétiques n'expliquent cependant qu'une infime partie des variations phénotypiques, suggérant que d'autres facteurs sont aussi à prendre en considération, tels que les interactions possibles entre variants génétiques et environnement.

Au cours de ma thèse, j'ai exploré les bases génétiques de certains phénotypes multifactoriels chez l'humain. Mon projet principal s'est articulé autour de la détection d'interactions éventuelles SNPs-environnement, susceptibles d'expliquer les variations quantitatives de lipides chez des patients de la Cohorte Suisse du VIH, les SNPs étudiés étant dans des gènes impliqués dans le métabolisme et le transport des lipides, et les facteurs environnementaux, les antirétroviraux prescrits à ces patients. Des approches statistiques de régression linéaire ont permis de modéliser les niveaux de lipides chez ces patients en fonction des SNPs, des médicaments et des interactions entre ces deux derniers. Dans le but de mieux cerner leur implication éventuelle dans l'apparition de dyslipidémie, réaction délétère sévère suite à la prise de traitements antirétroviraux, j'ai estimé le pouvoir prédictif de ces facteurs sur la réponse lipidique. Une validation croisée stratifiée par patient, n'a pas permis de valider l'hypothèse qu'une sélection "intelligente" de certaines interactions puisse avoir un meilleur pouvoir prédictif des niveaux de triglycérides qu'une sélection aléatoire d'interactions SNP-médicament. Néanmoins, tous les modèles linéaires testés, contenant les SNPs et/ou les médicaments, ont démontré un pouvoir prédictif sur la réponse lipidique, expliquant une fraction des variations des niveaux de lipides. Plus particulièrement, les SNPs sélectionnés lors de la validation croisée, ont permis de prédire en partie les niveaux de triglycérides. La dyslipidémie est un trait complexe dans lequel interviennent également d'autres facteurs non pris en compte lors de nos analyses. Les interactions étudiées n'ayant qu'un faible effet sur la réponse lipidique, les approches statistiques utilisées n'ont pas permis leur détection.

Les autres projets de ma thèse ont été réalisés dans le domaine des analyses d'association pan-génomique d'individus appartenant à la Cohorte de Lausanne. J'ai été impliquée à travers

des collaborations internationales dans la recherche et la détection de SNPs associées à des variations phénotypiques de certains traits, tels que les niveaux de calcium dans le sérum, l'indice de masse corporelle, les niveaux de glucose après glycémie provoquée, le syndrome métabolique et ses composantes. J'ai mis en pratique des méthodes de régression linéaire afin quantifier l'effet des SNPs. L'identification de variants génétiques, associés aux variations observées dans ces phénotypes pourra donner des perspectives d'applications cliniques majeures, telle que la prévention du risque et la médecine "personnalisée", sachant que ces phénotypes sont rattachés à un risque cardiovasculaire majeur chez l'humain.

Résumé grand public

Les variations de séquence au niveau de l'ADN humain peuvent influencer le risque de développer une maladie ou la réponse à un traitement. La recherche en génétique est en pleine expansion, générant une grande quantité de données génétiques qui combinées à des mesures phénotypiques tels que la taille, l'indice de masse corporelle, peuvent permettre de détecter des marqueurs génétiques, tels que les "Single Nucleotide Polymorsphisms" (SNPs), qui sont associés aux variations phénotypiques. Ces variants génétiques n'expliquent qu'une infime partie des variations phénotypiques, suggérant l'existence d'autres facteurs, tels que les interactions possibles entre variants génétiques et environnement.

Dans ce contexte là, je me suis intéressée aux variants génétiques au sein de gènes impliqués dans le métabolisme et transport des lipides (triglycérides, cholestérol) dans une population infectée par le virus du SIDA (la Cohorte Suisse du VIH). En effet, ces patients, suite au traitement par des médicaments antirétroviraux, peuvent développer des effets secondaires tels que des dyslipidémies, qui sont des variations anormales du taux de lipides dans le sérum. A l'aide de modèles statistiques basés sur la régression linéaire, j'ai évalué l'influence des variants génétiques, des médicaments, et plus particulièrement des interactions entre variants génétiques et médicaments sur la réponse lipidique. J'ai ainsi construit plusieurs modèles mettant en jeu ces différents facteurs et estimé le pouvoir prédictif de ces derniers sur les niveaux de triglycérides. Ces approches statistiques s'inscrivent dans une optique de médecine personnalisée où les patients pourraient dans le futur bénéficier de traitements antirétroviraux efficaces et engendrant le moins d'effets secondaires possibles.
J'ai également participé au travers de collaborations internationales impliquant de grandes cohortes d'individus, à l'analyse à grande échelle de données génétiques afin de révéler de potentielles associations entre des variations dans l'ADN et certaines maladies ou variations phénotypiques telles que les niveaux de calcium dans le sérum, les variations dans l'indice de masse corporelle, celles des niveaux de glucose après une glycémie provoquée, le syndrome métabolique et ses composantes. A l'aide d'outils informatiques et statistiques, j'ai analysé ces données et mis en évidence l'existence de variants génétiques associés à certains traits. A terme, ces études mèneront à une meilleure compréhension des bases génétiques de l'apparition de certaines maladies, et donc à une meilleure prise en charge des patients concernés.

Table of content

Abbreviations

ADR	Adverse Drug Reaction
AUC	Area Under the Curve
BMI	Body Mass Index
Chol	total cholesterol
DM2	Diabetes Mellitus type 2
EH	Essential Hypertension
FDA	Food and Drug Administration
FI	Fusion Inhibitor
FPR	False Positive Rate
GWAS	Genome-Wide Association Study
HAART	Highly Active AntiRetroviral Therapy
HDL	High Density Lipoprotein
HIV	Human Immunodeficiency Virus
IDL	Intermediate Density Lipoprotein
LDL	Low Density Lipoprotein
LR	Lipid Response
NHC	Non-HDL Cholesterol
N(t)RTI	Nucleoside (or nucleotide) Reverse Transcriptase Inhibitor
NNRTI	Non-Nucleoside Reverse Transcriptase Inhibitor
PI	Protease Inhibitor
ROC	Receiver Operating Characteristics
SHCS	Swiss HIV Cohort Study
SNP	Single Nucleotide Polymorphism
TG	Triglycerides
TPR	True Positive Rate
VLDL	Very Low Density Lipoprotein

1 General context

1.1 Computational genomics

Computational genomics refers to the use of computational analysis to decipher biological insights from genome sequences and related data, including both DNA and RNA sequences as well as other "post-genomic" data (i.e. experimental data obtained with technologies that require the genome sequence, such as genomic DNA microarrays [1, 2]). As such, computational genomics may be regarded as a branch of bioinformatics, but with a focus on using whole genomes (rather than individual genes) to understand the principles of how the DNA of a species controls its life cycle and response to the environment at the molecular level. With the current abundance of massive biological datasets, computational studies have become a very important means to biological discovery.

Genomic techniques have firmly established themselves as a standard tool in biological and biomedical research. Together with the rapid advancement of genome sequencing projects, microarrays and related high-throughput technologies have been key factors in the study of the more global aspects of cellular systems biology. While genomic sequence provides an inventory of parts, a proper organization and eventual understanding of these parts and their functions requires comprehensive views also of the regulatory relations between them. Genome-wide expression data offer such a global view by providing a simultaneous read-out of the mRNA levels of all (or many) genes of the genome.

The (human) post-genomic era began after the release of a rough draft of the human genome completed by the Human Genome Project in early 2001 [3, 4]. By 2007, the human sequence was declared "complete" (less than one error in 20,000 bases and all chromosomes assembled). The ensuing "genomic revolution" in biology has had a fundamental impact on the improvement of diagnosis, prevention and treatment of disease. Yet, while researchers have already started to use genomic data for predictive purposes in cancer research and clinical practice [5-8] , the next challenge lies in integrating the massive amount of data produced by different high-throughput technologies.

It has long been known that genetic variants influence the risk of developing certain diseases or determine certain traits. The recent generation of massive genotypic data using SNP arrays

[9, 10] for individuals with well-characterized phenotypic traits opened the field of genome-wide association studies (GWAS). These studies employ large cohorts of hundreds if not thousands of individuals to search for genetic differences (usually SNPs, but also copy number variations (CNVs)) that are correlated with phenotypes. Yet, for the large majority of traits the most significant genetic variants only explain a small fraction of the phenotypic variance, leaving room for other missing factors. In particular, the existence of SNP–environment interactions and their impact remains a difficult task within the context of large scale data analysis.

1.2 Pharmacogenectics and pharmacogenomics

The terms *pharmacogenetics* and *pharmacogenomics* tend to be used interchangeably, yet we would like to follow the common distinction that pharmacogenetics [11] is generally regarded as the study of genetically inherited variations in drug metabolism and response, while pharmacogenomics [12] is the general study of all of the many different genes that influence drug response. It is the broader application of genomic technologies to drug discovery and extended characterization of existing drugs. Pharmacogenetics usually considers one or at most a few genes of interest, while pharmacogenomics considers the entire genome or a large portion of it.

Nowadays, one of the main goals of pharmacogenetics is to optimize drug therapies by taking into account the patients' genotype (genetic makeup) and environment features (like diet or lifestyle) [13, 14]. Optimization means to aim for maximum efficacy of the treatment while minimizing adverse effects [15].

The promise of pharmacogenetics is to give rise to so-called *personalized medicine* [16, 17], where drugs will be administrated not only based on the phenotype, but also on the genotype of a patient. This should give rise to more powerful and safer medications, more accurate methods of determining appropriate drug dosages as well as advanced prophylactic screening for disease, all of which could potentially result in a decrease in the overall cost of health care.

The introduction of the microarray technologies [1, 2, 18], which enable scientists to gather genome-wide data on gene expression [19] or on single nucleotide polymorphisms (SNPs)

[9, 10] from many individuals, will play a major role in the future of both pharmacogenomics and pharmacogenetics [20, 21]. The explosion in both SNP and microarray data necessitated the development of a new means for cataloging and annotating these data (dbSNP, GEO) so that scientists can more easily access and use it for their research.

Yet, the analysis of the large datasets produced by these technologies still faces significant challenges. In particular, classical methods that were developed to analyze the effect of a single or a few SNPs cannot be easily extended to cope with large amounts of genomic data. Thus, new approaches are needed to efficiently and adequately deal with the large amounts of highly complex data generated by pharmacogenetic studies. Analysis and interpretation of these data will allow scientists to not only determine drug responses but also to study disease susceptibility and conduct basic research in population genetics.

1.3 Linear regression modeling and features selection

Linear regression is a well-established statistical tool used in genetics and epidemiological studies. It is a form of regression analysis in which observational data y are modeled by a linear function of the explanatory variable(s). While in simple regression y depends only on a single explanatory variable x, in multi-linear regression y is modeled as a linear combination of the explanatory variables $X = (x_1, x_2, \ldots, x_n)$ [22]. Thus, the most general linear model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n ,$$

where β_0 is referred to as the *intercept* and β_i ($i=1,\ldots,n$) are the *regression coefficients* or *effect sizes*.

Linear regression was the first type of regression analysis to be studied rigorously, and continues to be used extensively in practical applications. This is because linear models are easier to fit than models with non-linear terms and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications of linear regression fall into one of the following three broad categories:

- *Fitting*: Linear regression allows for accessing which fraction of the data for y can be *explained* in terms of a given set of explanatory variables $X = (x_1,x_2, ...,x_n)$ and what effect size can be attributed to each component x_i.
- *Prediction*: If the goal is forecasting, linear regression can be used to fit a predictive model to a set of observed data for y and X values. The fitted model can be used to make a prediction of the value of y for additionally collected data for X only.
- *Model selection*: Given a variable y and a number of features $X = (x_1,x_2, ...,x_n)$ that may or may not be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the x_j, and to assess which subsets of X contain the least redundant information about y.

The "least-squares method" is the standard fitting method used in linear regression. Here the best fit is characterized by parameters which minimize the sum of squared residuals, leading to unbiased and consistent estimators for β if the errors have zero mean, finite variance and are uncorrelated with the explanatory variables. The least-squares method corresponds to the maximum likelihood criterion if the experimental errors are normally distributed. Linear models may also be fitted in other ways, such as by minimizing a penalized version of the least squares loss function as in ridge regression [23]. Conversely, the least-squares approach can be used to fit models that are not linear models. Thus, while the terms "least-squares" and linear model are closely linked, they are not synonymous.

The linear modeling framework has various advantages: its mathematical formalisms and computational implementations are very well developed. Furthermore, its interpretation is straightforward. It directly models the mathematical relationship between parameters (such as SNPs, drug, patients' characteristics) and the phenotype of interest by examining the regression coefficients (effect sizes).

However, linear regression also suffers from various limitations: first, linear modeling makes assumptions about the nature of the data being modeled. In particular, it assumes a linear relationship, which is often not realistic. Second, tests of model parameters assume that the errors are normally distributed (linear regression is always unbiased). Finally, linear models

can suffer from overfitting depending on the complexity of the model and the amount of data [22].

Once a regression model has been constructed, it is important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include the fraction of explained variance R^2, analyses of the pattern of residuals to test for heteroskedasticity (non-constant variance of the errors) and hypothesis testing. Statistical significance of the overall fit can be assessed by an *F*-test; *t*-tests evaluate the significance of individual parameters.

The generalized linear model (GLM) is a flexible generalization of ordinary least squares regression [24]. It generalizes linear regression by allowing non-normal responses (via a link function**)**. Generalized linear models were formulated by John Nelder and Robert Wedderburn as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression. They proposed an iteratively re-weighted least squares method for maximum likelihood estimation of the model parameters. Maximum-likelihood estimation remains popular and is the default method for many statistical computing packages. Other approaches, including Bayesian approaches and least squares fits to variance stabilized responses, have been developed. Logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logit function (or "logistic curve"). It is a generalized linear model used for binomial regression. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical.

Model selection [22] is the task of selecting a mathematical model from a set of potential models, given the observed data. In the case of regression, it is also known as feature selection and it is the technique of selecting a subset of relevant features for building robust models. Feature selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset. The most popular form of feature selection is stepwise regression. It is a greedy algorithm that adds the best feature (or deletes the worst feature) at each round. The main control issue is deciding when to stop the algorithm. In statistics, there are a variety of optimality criteria that can be used for controlling feature selection. The oldest are Mallows'

Cp statistic and Akaike Information Criterion (AIC). Other criteria are Bayesian Information Criterion (BIC) which uses, minimum description length (MDL), Bonferroni / Risk Inflation, Criterion (RIC), and a variety of new criteria that are motivated by false discovery rate (FDR) [25].

2 HIV project

2.1 HIV/AIDS

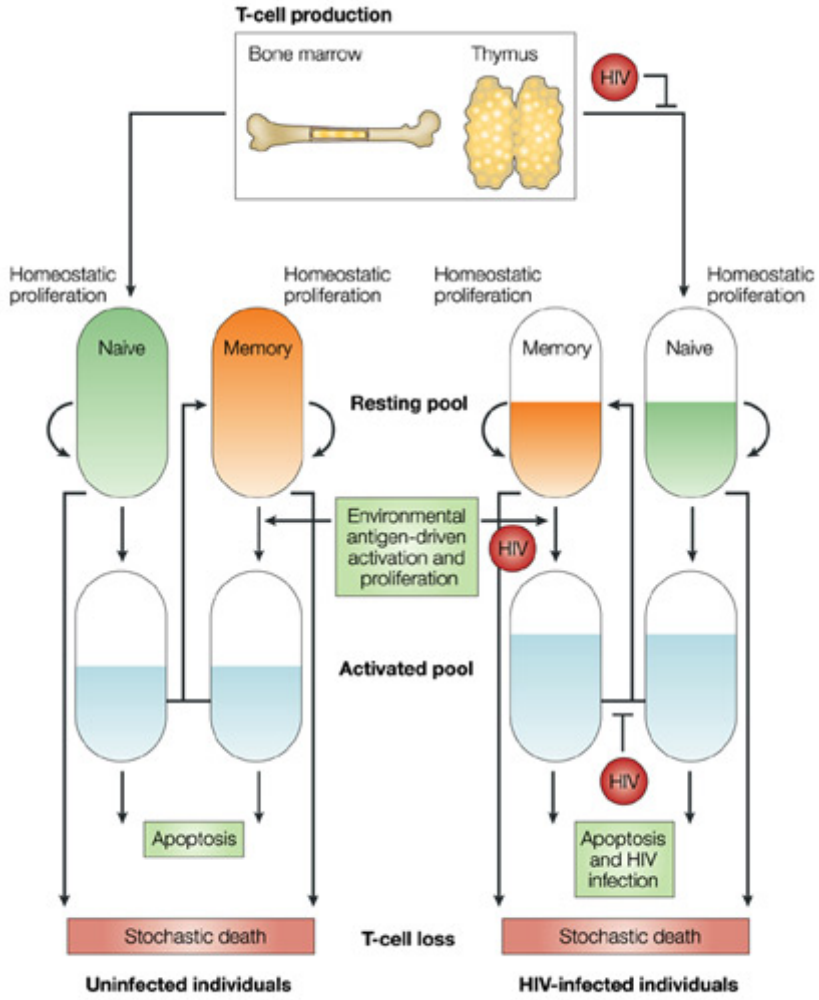
The Human Immunodeficiency Virus (HIV) is the virus that for the large majority of infected individuals leads to Acquired Immune Deficiency Syndrome (AIDS) [26, 27]. HIV belongs to a subset of retroviruses called lentiviruses (or slow viruses), which means that there is a rather long interval (sometimes years) between the initial infection and the onset of symptoms. Upon entering the bloodstream, HIV infects the CD4+ T cells and begins to replicate rapidly.

Scientists believe that when the virus enters the body, HIV begins to disable the body's immune system by taking advantage of its aggressive immune response to the virus to infect, replicate and kill immune system cells, such as helper T-cells, macrophages and dendritic cells. This process is facilitated by the recognition of the glycoprotein CD4 expressed on the surface of these cells. The immune system starts to fail when its cell levels (in particular of the CD4+ T-cells) fall below a critical threshold (~ 200 cells / mm^3). Gradual deterioration of immune function and eventual destruction of lymphoid and immunologic organs is central to triggering the immunosuppression that leads to AIDS and to giving rise to opportunistic infections (Fig. 1).

The term AIDS applies to the most advanced stages of HIV infection. There are four main stages in the progression of an HIV infected person developing AIDS. The period following the initial HIV infection is called the window period, during which HIV antibodies develop in the bloodstream. Seroconversion refers to the period of time during which the body is busy producing HIV antibodies, trying to protect itself against the virus. This is a highly infectious stage. The most common symptoms include fever, lymphadenopathy, pharyngitis, rash, myalgia, malaise, mouth and esophageal sores, and, less commonly, headache, nausea and vomiting, enlarged liver/spleen, weight loss, thrush, and neurological symptoms. After seroconversion, most people experience an asymptomatic period. This stage can last anywhere from six months to over ten years, varying from person to person. Although the person with HIV is experiencing no symptoms, the virus is still replicating inside the body and weakening the immune system. After this period, severe CD4+ T-cells loss leads to the symptomatic period, in which the body experiences the symptoms associated with HIV. This is the final stage before developing AIDS.

Most incidents of HIV infections occur through sexual contact. However, the virus can also be spread through blood transmission (e.g. by blood transfusions involving unscreened blood). It can also be transmitted from an HIV-infected mother to her child during pregnancy, birth or breastfeeding. There are a number of tests that are used to find out whether a person is infected with HIV. These include the HIV antibody test, P24 antigen test and polymerase chain reaction (PCR) test. Early detection offers more options for treatment and preventative care.

AIDS is pandemic and has killed more than 25 million people worldwide since 1981 (when the epidemic was first discovered). In 2005 alone, there were an estimated 4 million new HIV infections and 3 million deaths caused by AIDS (of which a third occurred in sub-Saharan Africa and more than 570 000 victims were children). Today, an estimated 40 million people (~0.6% of the world's population) are living with HIV [28], sub-Saharan Africa being the most seriously affected region with a prevalence of HIV infection of about 5%. AIDS ranks with malaria and tuberculosis as one of three most deadly infectious diseases among adults. HIV has orphaned more than 15 million children.



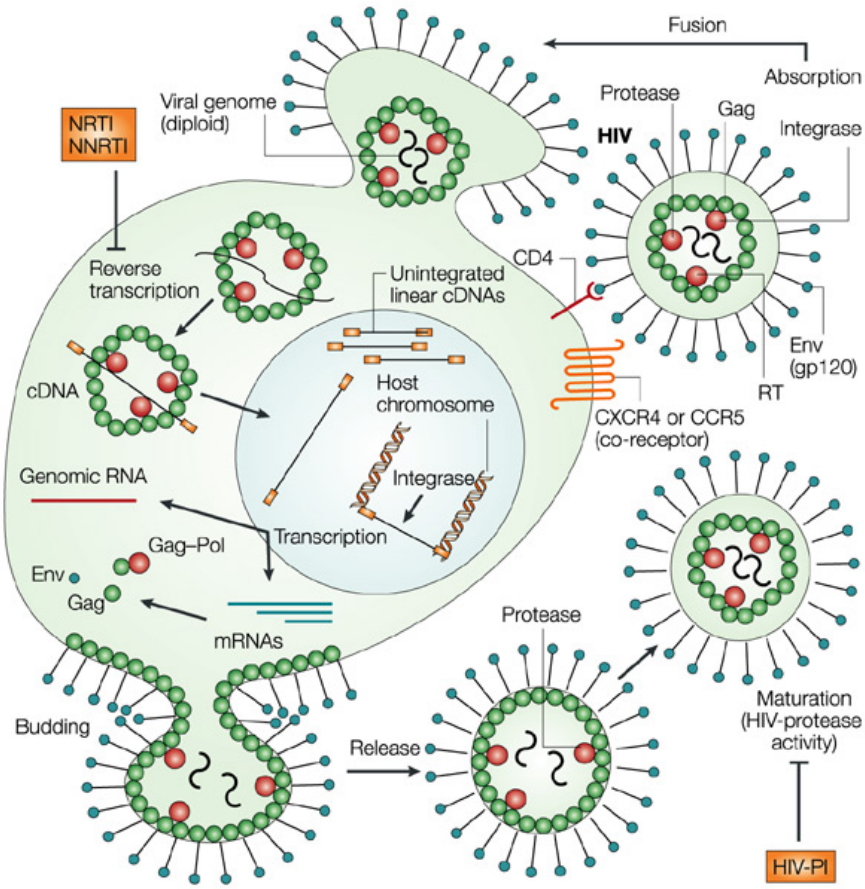
Modified from Monini et al., Nature Reviews Cancer, Nov. 2004

Figure 1: Normal immune system vs. HIV-infected immune system.

In a normal immune system (left side of figure), exposure to environmental antigens continuously activates naive and/or memory T cells, driving their replication and establishing and maintaining a pool of proliferating cells (activated pool). Activated cells become effectors and undergo apoptosis or survive to replenish the 'resting pool' as memory cells. Naive and memory T cells also die by 'natural' (stochastic) events — this is compensated by homeostatic proliferation. Production and maturation of T cells in the bone marrow and thymus also compensate for these losses. In patients infected with human immunodeficiency virus (HIV), however, several mechanisms perturb T-cell homeostasis (right side of figure). In the immune system of an individual infected with the HIV, the 'activated pool' of T cells is targeted by the virus, and this impairs the replenishment of the memory compartment. The input of T cells from the bone marrow and thymus is also reduced due to direct and indirect effects of HIV infection. But the main force that drives the immune system into collapse and that leads to overt immune deficiency in HIV infections is the chronic antigenic stimulation that results from uncontrolled HIV replication. As a result, the activated cell pool expands, whereas the pools of naive and the memory T cells progressively shrink.

2.2 HAART

There is no known cure for HIV. Only a vaccine is thought to be able to halt the pandemic, but HIV remains a difficult target, as there are a number of reasons why the development of an HIV vaccine is more challenging than that of other classic vaccines. Not only has HIV developed multiple mechanisms to avoid the immune response, but many different subtypes of the virus, known as clades, circulate in different regions of the world. Within those clades there is considerable variability. Based on the different challenges of an HIV vaccine design, it is believed that a candidate vaccine will provide robust protection against HIV infection only if it is devised to engage both cell-mediated immunity and antibody-based immune responses. For many years, there was no effective treatment for AIDS, but today, a number of drugs are available to treat HIV infection. In addition, there is progress with medications for better treatment of opportunistic infections and illnesses that affect people with AIDS. Since 1996, the best care for treating individuals infected with HIV is the so-called *Highly Active Anti-Retroviral Therapy* (HAART) [29]. It is expensive and therefore became the standard of care only in developed countries, while the worst affected regions in Africa cannot afford HAART for the entire affected population. HAART combines the administration of three or more drugs from at least two different classes. These drugs act at different levels of the viral replication cycle, preventing HIV from reproducing and destroying the body's immune system (Fig.2).



Nature Reviews | Cancer Modified from Monini et al., Nature Reviews Cancer, Nov. 2004

Figure 2: HIV replication cell cycle and HAART mode of action

Human immunodeficiency virus (HIV) infects several cell types, including CD4+ T cells, monocytes and dendritic cells. The infectious cycle begins with the adsorption of viral particles to the receptor CD4, which is present at the cell surface. This interaction, which is mediated by the HIV envelope (Env) protein gp120, leads to subsequent interaction of the gp120 V3 loop region with a co-receptor, usually a member of the seven-membrane-spanning chemokine-receptor families, the most important being CCR4 or CXCR5. After virus adsorption, the viral and cell membranes fuse together, and the viral 'core' (which includes the diploid viral genome) is released into the cytoplasm, where the virion-associated reverse transcriptase is activated and begins synthesizing viral cDNA. This is subsequently transported to the cell nucleus, where another virion-associated enzyme, the HIV integrase, catalyses the insertion (integration) of the viral cDNA into the host-cell genome. Transcription of the integrated viral cDNA leads to the production of genomic (unspliced) and messenger (spliced) RNA (mRNA) molecules that are transported to the cell cytoplasm. Translation of HIV mRNAs leads to the production of Env proteins and immature precursors of capsid (Gag) and viral polymerase (Pol) proteins. Immature Gag and fused Gag–Pol precursors are transported to the cell membrane, where viral progeny begin assembling and 'bud' from the infected cells. Viral particles released following budding, however, do not contain the characteristic HIV condensed core and are not infectious. Virus infectivity is acquired after particle maturation, which is mediated by the virion-associated HIV aspartyl protease. This enzyme cleaves the immature Gag and Gag–Pol precursors into functional polypeptides. These crucial steps of the HIV life cycle are targeted with nucleoside and non-nucleoside reverse-transcriptase (RT) inhibitors (NRTIs and NNRTIs, respectively), and with HIV-protease inhibitors (HIV-PIs).

The main classes of molecules used in HAART are:

- *Nucleoside (nucleotide) reverse transcriptase inhibitors (N(t)RTI):* they are incorporated into the viral DNA and prevent reverse transcriptase from adding nucleotides to form functional DNA. They include abacavir, didanosine, emtricitabine, lamivudine, stavudine, tenofovir, and zidovudine.
- *Non-nucleoside reverse transcriptase inhibitors (NNRTI)*: they attach themselves to reverse transcriptase to prevent HIV from converting RNA into DNA, thus preventing the cell from producing new virus. They include delavirdine, efavirinz, etravirine, and nevirapine.
- *Protease inhibitors (PI):* they attack the HIV enzyme protease and include amprenavir, atazanavir, darunavir, fosamprenavir, indinavir, lopinavir, nelfinavir, ritonavir, and tipranavir.
- *Entry inhibitors:* They prevent HIV from entering healthy CD4 cells by targeting the CCR5 protein. Maraviroc and Selzentry are Food and Drug administration (FDA)-approved entry inhibitors.
- *Fusion inhibitors (FI)*: they stop the virus from entering cells by targeting the gp41 protein on HIV's surface. Enfuvirtide is a FDA-approved fusion inhibitor.
- *Integrase inhibitor*: they block the action of an enzyme produced by HIV that allows it to integrate into the DNA. It is effective against HIV that has become resistant to other antiretroviral drugs. Raltegravir is the only FDA-approved integrase inhibitor, but several other integrase inhibitors are under clinical trials. HIV integrase inhibitors may be useful not only for inhibiting HIV but also for treating dyslipidemia induced by chronic HIV protease inhibitor therapy.

These drugs can reduce the amount of HIV in the bloodstream to very low levels and often enable the body's immune cells to rise back to normal levels. Indeed a dedicated combinatorial treatment under regular and systematic supervision and adjustment has considerably improved the life expectancy and quality of HIV infected individuals over recent years in the developed world [29, 30].

An important factor reducing morbidity due to AIDS has been the emergence of efficient drugs for fighting opportunistic diseases (caused by bacterial, viral, fungal or protozoan pathogens) in immuno-compromised patients. Specifically, there now exist several drugs that help to prevent pneumocystis, carinii pneumonia, toxoplasmosis, cryptococcus, and

cytomegalovirus infection. Once opportunistic infections have occurred, the same drugs can be used at higher doses to treat these infections. Moreover, AIDS related morbidity has also been reduced as a result of improved chemotherapy for treating cancers such as lymphomas or Kaposi's sarcoma that have higher incident rates in people suffering from AIDS.

Ongoing research on anti HIV therapies currently focuses on decreasing side effects of the present drugs and on the further improvement of drug regimens. The latter is targeted at better adherence, determining the optimal sequence of regimens to manage drug resistance, and identifying new targets for anti-HIV medications [31]. There is also ongoing research studying ways of restoring immune systems damaged by HIV.

Adverse effects of antiretroviral drugs are well known and are a major challenge for physicians prescribing a particular therapy. They vary by drug, by ethnicity, by individual, and by interaction with other medications or the patients' diet, including their alcohol consumption [31]. The most common adverse effects experienced by patients taking antiretroviral drugs are: diarrhea, nausea, headache, fatigue, dyslipidemia, lipodystrophy, abdominal pain, liver and renal failure, kidney diseases, pancreatitis, insulin resistance, increase of cardiovascular risks, osteoporosis, anemia, neutropenia, insomnia, peripheral neuropathies, and birth defects [32, 33].

2.3 Pharmacogenetics applied to HIV treated population and personalized medicine

Currently, there is no simple way to determine whether people will respond well, poorly, or not at all to medication. Therefore, pharmaceutical companies tend to develop general therapeutic strategies based on the average patient response, rather than "customized" treatments. Yet, the limitation of this approach is becoming more and more obvious and costly in light of the wide range of serious adverse drug reactions (ADRs) mentioned above. The detection, assessment, understanding and prevention of ADRs are obviously not only a concern for antiviral medications but it is a field of itself known as *pharmacovigilance*. In a study published in 1998 [34], it was estimated that in 1994 more than 2.2 million hospitalized patients had serious ADRs and that an estimated 76,000-137,000 had fatal ADRs. The latter implies that ADRs range between the fourth and sixth leading cause of death in the United States. Since then, the Research on Adverse Drug events And Reports (RADAR) has been

initiated. Its goal is to describe previously unrecognized, serious ADRs and identify new patient populations at high risk for previously identified serious ADRs [35].

In addition to their impact on human health, ADRs also have a significant impact on healthcare costs [36]. ADRs are common causes of hospitalization and lead to large costs to society. The cost of hospitalization is, however, only a part of the total cost, as most ADRs never come to clinical attention. There are two main costs associated with ADRs, the cost of treating illnesses due to ADRs and the cost of avoiding them [37].

ADRs can be divided into two broad clinical categories, about 80% are pharmacological (or of type A) and 20% are idiosyncratic (type B). ADRs of type A are generally dose-dependent, related to the pharmacokinetic properties of a drug (absorption, distribution, metabolism and excretion), and resolve when the dose is reduced. The type B ADRs are more enigmatic an usually cannot be explained by the known pharmacology of the drug [38]. Factors that contribute to type B ADRs of a drug are not yet completely understood, but are likely to include environmental factors such as diseases, alcohol, smoking and diet. Genetic factors can also affect the susceptibility to both types of ADRs. Genetic polymorphisms are a source of variation of drug response in the human body [39]. Most interest has centered on the involvement of pharmacokinetic factors and in particular drug metabolism. Several polymorphisms were discovered that affect gene-expression encoding in phase I and phase II metabolic pathways [40, 41], such as cytochrome P450 enzyme, thiopurine methytransferase, UDP-glucuronosyl transferase 1A. Membrane transporters also play an important role in terms of drug pharmacokinetics. Variations in the multi-drug resistance gene *ABCB1/MDR1* increase the risk of drug toxicity. In addition to the genetic polymorphisms in the pharmacokinetics factors, there are also genetic variations in pharmacodynamic factors (drug targets) such as variations in enzyme structures, variations in receptors of a certain drug, mutations in ion channels, and variations in immune response genes (*HLA* types) [39].

Thus, genotypic profiling of patients has the great potential to allow for more customized treatment [42]. In particular, many therapeutically active compounds that never entered the market because they failed in clinical trails may find their way back to commercialization if their ADRs can be demonstrated to occur only for individuals of a certain genotype. Similarly, drugs that were not found to be sufficiently active in the general population, may still re-enter the market if they can be proven to be highly active for a significant

subpopulation that differs in one or more genetic markers. Clinical trials restricting its subjects to those with a favorable genotype could be smaller, faster, and therefore less expensive, leading to reduced drug costs. If successful, such genetically stratified medications could potentially increase both physicians' and patients' confidence in prescribing and taking a drug.

As an example, Warfarin is used clinically as an anticoagulant but requires periodic monitoring and is associated with adverse outcomes. Recently, genetic variants in the gene encoding Cytochrome P450 enzyme CYP2C9, which metabolizes Warfarin, and the vitamin K epoxide reductase gene (*VKORC1*), a target of coumarins, have led to commercially-available testing that enables more accurate dosing based on algorithms that take into account the age, gender, weight, and genotype of an individual. Targeted therapy is the use of medications designed to target aberrant molecular pathways in a subset of patients with a given cancer type. For example, Herceptin is used in the treatment of women with breast cancer in which HER2 protein is overexpressed [7]. Tyrosine kinase inhibitors such as Gleevec have been developed to treat chronic myeloid leukemia (CML), in which the *BCR-ABL* fusion gene is present in >95% of cases and produces hyperactivated abelson-driven protein signaling [43]. These medications specifically inhibit the Abelson tyrosine kinase (*ABL*) protein and are thus a prime example of "rational drug design" based on knowledge of disease pathophysiology. As another example, the aim of the five-year, international drug-sensitivity study is to find the best combinations of treatments for a wide range of cancer types: roughly 1000 cancer cell lines are exposed to 400 anticancer treatments, alone or in combination, to determine the most effective drug or combination of drugs in the lab. The therapies include known anticancer drugs as well as others in pre-clinical development. The first data release confirmed several genes that predict therapeutic response in different cancer types. These include sensitivity of melanoma, a deadly form of skin cancer, with activating mutations in the gene *BRAF* to molecular therapeutics targeting this protein, a therapeutic strategy that is currently being exploited in the clinical setting.

In the context of an HIV population, pharmacogenomics is the study of the genetic basis for abnormal drug reactions of patients under antiretroviral therapies. The introduction of HAART and its use has considerably enhanced life expectancy of HIV/AIDS individuals, reducing viral replication into the host cells. Nevertheless, being a long-term treatment, HAART induces toxicities and drug resistance due to the high viral genetic variability

(replication cycle and mutation rate). Drug metabolism and toxicity can vary between individuals leading to different degrees of treatment efficacy and toxicity effects. The reason for HAART failure is multi-factorial, including drug adherence, virological, immunological and pharmacological factors. Yet, it is also likely that genetic variations of HIV infected individuals account for a major part of the variability in antiviral drug responses [44]. Genetic studies and technologies (such as GWAS) are generating new hypotheses in terms of potential links between genotypes and phenotypes giving insights on which host genetic polymorphisms could influence the HAART drug response, i.e. pharmacokinetics and pharmacodynamics, hypersensitivity reaction syndromes, hepatotoxicity, central nervous system side effects, hyperbilirubinemia, peripheral neuropathy, lipodystrophy, hyperlipidemia, pancreatitis and renal toxicity [31]. To date, pharmacogenetics studies of antiretroviral drugs have identified several genetic polymorphisms in critical metabolizing and drug-transporter genes which can influence specific antiretroviral drug pharmacokinetics (Table 1). As examples, some polymorphisms in the *ABC* genes, which encode proteins responsible for carrying many types of drugs across the membranes, combined with PIs and NRTIs treatment showed associations with respectively increase risk of hyperlipidemia, unconjugated hyperbilirubinemia and jaundice, and higher intracellular exposure of the triphosphate metabolite. PIs and NNRTIs drugs in combination with genetic variations in the cytochrome P450 enzyme gene, responsible for oxidative metabolism of the majority of the drugs, show respectively higher drug exposure, faster oral clearance, and greater plasma exposure [31] (Table 1).

The ultimate goal of pharmacogenomics is to transfer the research knowledge into clinical use. The promise is that eventually a "personalized HAART" will lead to both a maximization of virological efficacy and a minimization of ADRs of the medications in a cost-efficient manner. Yet, developing such an individual HIV treatment may still take many years. Researchers are still at the beginning of developing powerful methods for analyzing the massive genetic data, facing issues like minimizing false discoveries, dealing with small study size (resulting in inadequate statistical power) as well as sometimes unavoidable selection bias or ethnic bias [31].

Drug, drug class	Gene, alllele(s)/polymorphism(s)	Reported associations	Additional findings and comments
Abacavir	HLA-B*5701	Increased risk of hypersensitivity reaction	Pharmacogenetic testing shown to be cost-effective. Pharmacogenetic testing before abacavir prescription recommended by all guidelines
TDF	ABCC2 (MRP2) 1249G>A	Increased risk of renal proximal tubulopathy	To be confirmed in other populations
3TC, ZDV	ABCC4 (MRP4) 3724G>A, 4131T>G	Higher intracellular exposure of the triphosphate metabolite	Uncertain clinical significance
NRTIs	TNFα 238G>A	Earlier onset of lipoatrophy	Negative findings reported by some authors
NRTIs	Mitochondrial DNA (haplogroup T)	Increased risk of peripheral neuropathy in some reports	Tissue-specific mitochondrial DNA depletion may also play a role in NRTI toxicity
NRTIs	HFE 845G>A	Reduced risk of peripheral neuropathy	Negative findings reported by some authors
NRTIs	CFTR 1717-1G>A, IVS8 5T SPINK-1 112C>T	Increased risk of pancreatitis	Reported also in the general population
NVP	HLA-DRB1*0101	Increased risk of hypersensitivity reaction and hepatotoxicity	CD4 cell percentage greater then 25% associated with increased risk
NVP	HLA-cw8	Increased risk of hypersensitivity reaction in some populations	
NVP, EFV	ABCB1 (MDR1) 3435C>T	Reduced risk of hepatotoxicity	
EFV	CYP2B6 516G>T, 983T>C	Greater plasma exposure and increased risk of CNS side effects	Reports of successful EFV dose individualization
EFV NVP	CYP2B6 516G>T, 983T>C	Greater plasma exposure	To be confirmed in other populations
EFV	MDR1 3435C>T	Reduced plasma exposure	Negative findings reported by some Authors
EFV	MDR1 3435C>T	Increase in HDL-cholesterol	To be confirmed in other populations
ATV, IDV	UGT1A1*28	Unconjugated hyperbilirubinemia and jaundice	
ATV	ABCB1 (MDR1) 3435C>T	Unconjugated hyperbilirubinemia and jaundice	Greater plasma levels
NFV	CYP2C19*2 (681G>A)	Higher drug exposure	To be confirmed in other populations
IDV	CYP3A5*3 (A6986G)	Faster oral clearance	To be confirmed in other populations
PIs	APOA5 –1131T>C, 64G>C	Increased risk of hyperlipidaemia	
PIs	APOC3 482C>T, 455C>T, 3238C>G	Increased risk of hyperlipidaemia	
PIs	APOE ε2 and ε3 haplotypes	Increased risk of hyperlipidaemia	
PIs	ABCA1 2962A>G CETP 279A>G	Increased risk of hyperlipidaemia	
RAL	UGT1A1*28/*28	Modestly higher plasma levels	Clinically not significant
MVC	CCR5 WT/Δ32	No effect on virological response	

ABC genes: genes in the ATP-binding cassette (ABC) family that encode for transporter proteins responsible for carrying many types of drugs across cell membranes.
APO (apolipoproteins): lipid-binding proteins, divided in five major classes (A, B, C, D, E) and several sub-classes, are the constituents of the plasma lipoproteins.
CCR5 (chemokine receptor 5) gene: located on chromosome 3, encodes for the CCR5 protein, the chemokine receptor for the chemokines RANTES, MIP-1alpha and MIP-1beta.
CETP (cholesteryl ester transfer protein): plasma protein that facilitates the transport of cholesteryl esters and triglycerides between the lipoproteins.
CFTR (cystic fibrosis transmembrane conductance regulator): mutations in this gene are involved in a number of clinical conditions including cystic fibrosis, male infertility and idiopathic pancreatitis.
SPINK-1 (serine protease inhibitor, Kazal type 1): encodes for a trypsin inhibitor in the cytoplasm of pancreatic acinar cells.
CYP (cytochrome P450): a superfamily of heme-binding proteins responsible for oxidative metabolism of the majority of drugs.
HFE (hemochromatosis): hereditary hemochromatosis is a multisystem iron overload disorder cases due to mutations that in the HFE gene resulting in altered iron adsorption and transport.
HLA (human leukocyte antigen): group of genes resides on chromosome 6 that encodes cell-surface antigen-presenting proteins and many other genes.
MDR1 (multidrug-resistance 1) gene: gene (also called ABCB1) that encodes P-glycoprotein, the multidrug efflux pump transporter that eliminates many drugs from cells and tissues.
TNF (tumor necrosis factor): a cytokine involved in systemic inflammation and acute phase reactions.
UGT (uridine diphosphate-glucuronosyltransferase): a class of enzymes including UGT2B7, UGT1, and UGT1A1, an enzyme of the glucuronidation pathway that transforms small lipophilic molecules, such as steroids, bilirubin, hormones, and drugs, into water-soluble, excretable metabolites.

Table 1: Summary of most relevant (established and putative) genetic determinants of antiretroviral drug pharmacokinetics and toxicity. From Tozzi V, antiviral research, 2009.

2.4 HAART induced dyslipidemia and cardiovascular risk

The term *dyslipidemia* refers to an abnormality in the amount of lipids or lipoproteins (VLDL, LDL, HDL, IDL) in the blood. The most common types of dyslipidemia are high cholesterol, called "hypercholesterolemia" or "hyperlipoproteinemia"; and high triglycerides named "hypertriglyceridemia". When high cholesterol (LDL) and triglycerides occur together, the condition is called "combined hyperlipidemia", which is the most common form of dyslipidemia in societies of developed countries. Low levels of cholesterol may also occur, called "hypocholesterolemia" or "hypolipoproteinemia", usually for genetic reasons. The prolonged elevation of insulin levels can lead to dyslipidemia. The increased levels of O-GlcNAc transferase (OGT) are known to cause dyslipidemia [15]. The treatment for dyslipidemia depends on the condition, but usually involves maintaining a healthy diet, exercise and taking medications, with lipid lowering agents (LLA), usually statins or fibrates.

For HIV patients, dyslipidemia is often a result of their antiviral therapies [45]. HAART can lead to elevations in serum levels of total cholesterol and triglycerides, reduction in HDL cholesterol, alterations in the distribution of body fat (lipodystrophy), as well as increase in insulin resistance and diabetes, which are major risk factors for cardiovascular diseases

(CVDs) [46, 47]. PIs and NRTI appear to be involved, through direct metabolic effects and an indirect effect of PIs and NRTI-related lipodystrophy [47]. Dyslipidemia occurs in up 70%-80% of HIV-infected individuals receiving HAART and can be associated with all available PIs, although hypertriglyceridemia appears to be more frequent in patients treated with ritonavir, saquinavir/ritonavir or lopinavir/ritonavir [48]. HIV protease inhibitor therapy induces endoplasmic reticulum (ER) stress, activating the unfolded protein response, which is an important signaling pathway in HIV protease inhibitor-induced metabolic syndromes. In addition, HAART itself causes metabolic syndrome in a high proportion of patients, characterized by lipodystrophy/lipoatrophy, dyslipidemia and insulin resistance that may be associated with an increase risk for coronary artery disease and stroke [48, 49]. In order to reduce the incidence of adverse metabolic effects, many studies have tried to assess the risk for CVDs by evaluating dyslipidemia in HIV treated patients, trying to establish the best combinations of drugs [48]. Other options for the treatment of lipid disorders include rosuvastatin, ezetimibe and fish oil and the use of HIV integrase inhibitors because the lipid goals of patients are not always achieved by the therapy recommended in the current lipid guidelines [50].

More specifically, host genetic variants (SNPs) were reported in the literature to be associated with plasma lipid level variations in response to drug therapy [31]. Polymorphisms in apolipoproteins (*APO*) are associated with hyperlipidemia and cardiovascular events in the general population. *APO* polymorphisms have been extensively studied regarding PI-associated metabolic and morphological abnormalities[51]. In the general population, polymorphisms in *APOC3* and *APOE* are associated with hyperlipidemia and several research groups showed the association of *APOC3, APOE, APOA1, APOA5* and *TNF-alpha* polymorphisms with the development of dyslipidemia and lipodystrophy. In addition, SNPs of *ABCA1, APOA5, APOC3, APOE,* and *CETP* genes contributed to plasma triglyceride and LDL–cholesterol levels during HAART [31, 52] (Table 1). Hyperlipidemia has been associated not only with PIs but also with an NNRTI drug, Efavirenz. EFV plasma levels are influenced by a polymorphism in the *ABCB1/MDR1* gene, causing an increase in HDL-cholesterol [53]. Finally, the contribution of 42 SNPs, most of them associated with lipid disorders in the general population, has been validated in the Swiss HIV Cohort Study (SHCS). The study also includes environmental factors, in particular the therapy regimens and estimates their impact on lipid disorders. As a result, the SNPs and drugs cumulative influence has been highlighted to contribute to dyslipidemia in HIV treated individuals [54].

Even if the HAART associated metabolic and morphological abnormalities are multi-factorial, next-generation HAART could increasingly benefit from genotype-guided drug choice, towards the ultimate goal of a personalized therapy (Fig. 3).

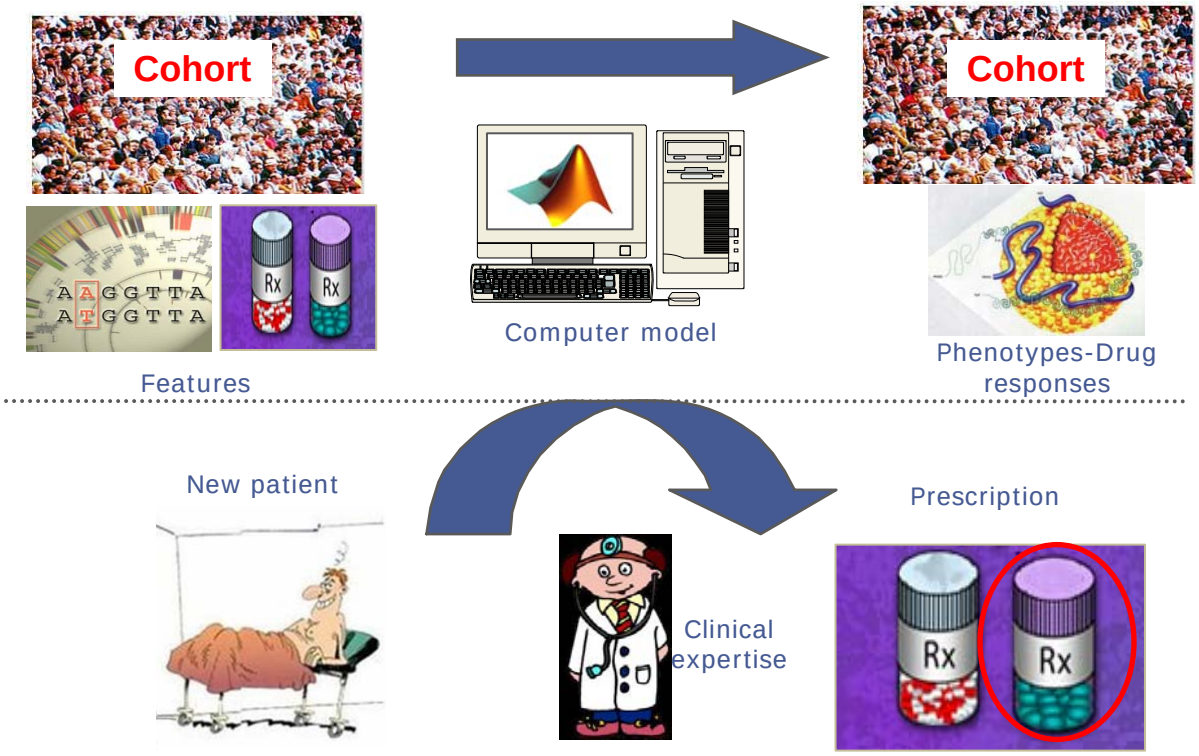


Figure 3: Goals of pharmacogenetics in the context of the Swiss HIV cohort data.
Different types of data (the patients' characteristics, the genotypic information, the therapies and the phenotype of interest (in this case the lipid levels) are represented in this schematic view. The goal is to combine these data into a computer model able to explain the variations we observe in the phenotypes. Ultimately such a model should allow for predicting the drug response of a new patient suffering from AIDS according to his or her features, which could be beneficial in terms of drug prescription.

2.5 Goals of the project

The main goals of my HIV project were the following:

1. Better understanding of the structure of the genetic and phenotypic datasets generated by the Swiss HIV cohort study (see http://www.shcs.ch/ and description below).
2. Exploration of the relationship between patients' drug responses, patients' features (including SNPs), and therapy characteristics. We were particularly interested in the potential implications of SNP-drug interactions in the understanding of variations observed in HIV patients lipid levels.
3. Evaluation of the predictive performances of different lipid models using genotypes, treatment features and their interactions.

2.6 Data, Methods and Results

2.6.1 The Swiss HIV Cohort Study

2.6.1.1 General description

Three datasets (Lipogen 2a, Lipogen 2b, and Lipogen 3) were provided by the Infectious Diseases Service of the Centre Hospitalier Universitaire Vaudois (CHUV). The data were gathered within the framework of the Swiss HIV cohort study. All three Lipogen datasets focus on longitudinal lipid measurements from HIV treated patients as phenotypes. Lipogen 2b extended Lipogen 2a mainly by a significantly longer follow-up period, while Lipogen 3 also included more individuals and a much larger set of candidate genes.

All patients were monitored during their treatment for phenotypes indicating obesity, metabolic syndrome and diabetes, including anthropometric variations (changes in hip and waist circumference, weight). To assess the presence of dyslipidemia, lipid determinations included the routine assessments of total cholesterol triglycerides (Chol), triglycerides (TG) and high density lipoprotein (HDL). The fasting state (defined as more than eight hours without caloric intake) was recorded for all blood draws [52]. Although fasting was requested, the compliance was only ~20%. Additional information includes the records of the antiviral drug combination, and patients' characteristics, including genotypes for candidate genes involved in lipid metabolism. Thus, conceptually, each dataset can be divided into the response variables (i.e. the lipid measurements) and the explanatory variables, where the latter is assumed to be causal with regard to the former. Specifically the explanatory variables can be subdivided as follows:

- **Patients' characteristics**: many of these features (like sex, height, ethnicity and presumed mode of HIV transmission) are fixed, but some (like age, body mass index (BMI), weight, and fasting state) may vary longitudinally from one measurement to the other.
- **Genotypic data**: Single Nucleotide Polymorphisms (SNPs) in candidate or validated genes involved in lipid transport and metabolism were genotyped for each patient.
- **Treatment data**: several drugs were used in highly active antiretroviral therapy (HAART) based on viremia and CD4+ T-cells count (see sections 2.1 & 2.2 for details). In general, the treatment varied longitudinally (meaning that each patient's treatment varied over time). We only considered the treatment regimen immediately prior to a measurement as causal

(ignoring the long-term history of treatments, which is usually assumed to be negligible for HIV patients).

As for the **response data** we considered not only the directly measured Chol, TG and HDL, but also non-HDL cholesterol (NHC) that was calculated as NHC = Chol – HDL. Indeed, as we mentioned in section 2.4, antiretroviral drugs can elevate the level of lipids in the blood, especially triglycerides and total cholesterol. This type of variation may cause dyslipidemia which has been associated with many illnesses, most notably cardiovascular diseases.

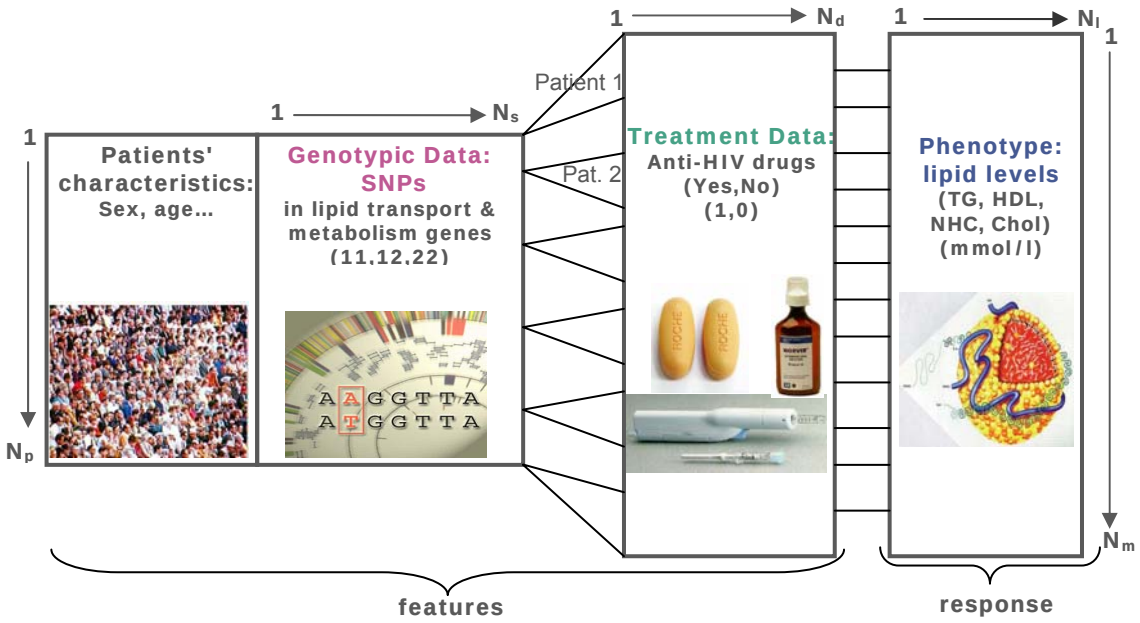


Figure 4: Structure of the Swiss HIV cohort data.

The dataset can be subdivided in four subsets: the patients' characteristics, the genotypic information, the therapies and the phenotype of interest (in this case the lipid levels). The dimensions of the measurements changed over time (longer follow up and new dataset). N_p: number of patients; N_s: number of SNPs; N_d: number of drugs; N_l: number of lipids; N_m: number of measurements

2.6.1.2 Specificities of each dataset

The dataset that was originally used included 4717 lipid measurements from 438 treated HIV patients, out of which ~75% were male. On average, each patient had been examined about 10 times over a follow-up period of four years. 20 SNPs from 12 candidate genes were genotyped. Patients were treated with a subset of a total of 18 antiviral drugs. Also many patient characteristics were recorded or computed for each patient such as age, sex, fasting state, diabetes status, body mass index (BMI), waist circumference, lipid lowering agents (LLA), smoking, CD4+ cell count, viral RNA levels, ethnicity, as well as risk behavior that may have led to the HIV infection. Hereafter, I will refer to this dataset as Lipogen 2a. I mainly used this dataset for my initial analyses described in sections 2.6.2.1 and 2.6.2.2.

Subsequently a further 1466 new lipid measures were included in the Lipogen 2 study. Altogether, the extended dataset contains a total of 6183 measurement from the same 438 patients, recorded during a follow up of six years. We removed four out of 20 SNPs (adrb2-27e, tnfa-238a, tnfa-308a and mtp-493t) because they had a high rate of missing genotype calls (20% on average). We also discarded three out of the 18 drugs from the dataset (emtricitabine (ETC), tipranavir (TPV), enfuvirtide (T20)) because they were very rarely given to any of the patients. Hereafter I will refer to this extended dataset as Lipogen 2b. Work described in sections 2.6.2.3-5 focused on this dataset.

Most of our analyses (sections 2.6.2.6-8) were performed on a significantly extended dataset from the Lipogen 3 study. It contains 12,170 lipid measures from 752 HIV patients. These 752 subjects include: 418 individuals from Lipogen 2 [52], 123 subjects who developed Diabetes Mellitus Type 2 (DM2) [30] and 211 additional subjects selected from the SHCS with similar characteristics as Lipogen 2 patients. The main patients' characteristics are shown in Table 2.

Characteristic	Study Participants (n =752)
Age (years)	47.36 +/- 10.43
Men/Women (%)	76 / 24
Fasting state (%)	21.5
LLA (%)	13.9
Log_{10} CD4+ T-cell count	2.66 +/- 0.25
Log_{10} viral RNA	1.83 +/- 1.13
Body Mass Index (BMI)	23.67 +/- 3.8
Waist circumference (cm)	87.64 +/- 11.54
Caucasians (%)	85.9
Diabetes Mellitus type 2 (DM2) (%)	16.8
Smokers (%)	31.5
Median number of lipid measurements/patient	16
Median follow-up period (years) (2000-2008)	7.8
TG (mmol/L)	2.52 +/- 2.59
HDL (mmol/L)	1.253 +/- 0.449
NHC (mmol/L)	4.01 +/- 1.33
Chol (mmol/L)	5.29 +/- 1.40

Table 2: Characteristics of the Lipogen 3 patients.

The values are shown in terms of mean +/- standard deviation or in the unit specified in brackets.

The Lipogen 3 dataset included a much larger set of a total of 58 SNPs. Their selection was mostly driven by genome-wide association studies on dyslipidemia (HDL: 20 SNPs; non-HDL/NHC: 14 SNPs; TG: 22 SNPs), new-onset diabetes mellitus type 2 (12 SNPs), metabolic syndrome (42 SNPs), and anthropometric measurements (one SNP) in the 752 HIV infected subjects followed in the SHCS. All these SNPs were near or in genes coding proteins involved in pathways relevant to beta cell function, or lipid/glucose metabolism/transport (Table 3 and Appendix 1).

TRIGLYCERIDES							
rs number	Nearest Gene	SNP type	Alleles	Chr	HWE	MAF	% NaNs
rs780094	*GCKR*	intronic	C/T	2	0,657	0,438	0
rs429358	*APOE*	exonic	T/C	19	0,029	0,114	0
rs7412	*APOE*	exonic	C/T	19	0,418	0,072	0
rs693	*APOB*	exonic	G/A	2	0,767	0,439	0
rs708272	*CETP*	intronic	G/A	16	0,878	0,390	0,600
rs328	*LPL*	exonic	C/G	8	0,142	0,113	0,386
rs2197089	*LPL*	3' down	A/G	8	0,262	0,425	0,477
rs6586891	*LPL*	intergenic	A/C	8	0,179	0,319	0,468
rs4775041	*LIPC*	intergenic	G/C	15	0,113	0,243	0,600
rs3135506	*APOA5*	exonic	G/C	11	0,550	0,066	0,600
rs2854117	*APOC3*	5' up	C/T	11	0,794	0,299	0,600
rs2854116	*APOC3*	5' up	T/C	11	0,940	0,402	0
rs5128	*APOC3*	3' down	C/G	11	0,557	0,105	0,468
rs662799	*APOA5*	5' up	A/G	11	0,809	0,082	0
rs16996148	*CILP2*	3' down	G/T	19	0,122	0,092	0,090
rs4846914	*GALNT2*	intronic	A/G	1	0,159	0,444	0,509
rs17145738	*TBL2*	3' down	C/T	7	0,705	0,109	0,477
rs17321515	*TRIB1*	3' down	A/G	8	0,883	0,471	0,337
rs12130333	*ANGPTL3,DOCK7,ATG4C*	intergenic	C/T	1	0,471	0,186	0
rs1748195	*ANGPTL3,DOCK7*	intronic	C/G	1	0,038	0,328	0,477

Table 3: List of SNPs in Lipogen 3 previously associated to TG variations.

The table contains for each SNP, the rs number, the type, the nearest gene, the alleles, the chromosomal location (Chr), the Hardy-Weinberg Equilibrium (HWE) test *p*-value, the minor allele frequency (MAF) and the percentage of missing genotypic data (NaNs).

Among the 32 drugs of the Lipogen 3 dataset, we only used half of them (16) in our analyses, since the remaining drugs were given less than 1% of the time (Table 4). Patients were treated with a combination of these drugs, usually including at least one drug from the three major classes of medications (NRTI, NtRTI and PI), acting at different levels of the virus life cycle.

Drug	Full Name	% given	Type
3TC	Lamivudine	66,47	NRTI
ABC	Abacavir	26,54	NRTI
AZT	Zidovudine	46,98	NRTI
D4T	Stavudine	16,26	NRTI
DDI	Didanosine	13,67	NRTI
ETC	Emtricitabine	3,87	NRTI
TNV	Tenofovir	16,64	NtRTI
APV	Amprenavir	1,38	PI
RTV	Ritonavir	15	PI
LPV	Lopinavir	12,51	PI
ATV	Atazanavir	7,32	PI
SQVH	Saquinavir hard gel	5,14	PI
NFV	Nelfinavir	16,63	PI
IDV	Indinavir	4,23	PI
NVP	Nevirapine	8,6	NNRTI
EFV	Efavirenz	25,73	NNRTI

Table 4: List of drugs used in HAART in Lipogen 3.
The table contains the name and full name of each drug, the percentage given within the SHCS and the drug type: NRTI (nucleoside reverse transcriptase inhibitor), NtRTI (nucleotide reverse transcriptase inhibitor), PI(protease inhibitor), NNRTI (non-nucleoside reverse transcriptase inhibitor).

2.6.2 Computational analyses and results

All analyses were performed in Matlab® under Windows or Linux operating systems. For developing my analysis approaches, I focused on TG as the phenotypic response.

2.6.2.1 Structure of the genetic and phenotypic datasets

Our first goal was to study the structure of the different data subsets and to identify and remove potential problems in these data. All lipid levels were $\log_{10}$-transformed (because the distributions of the log-values were closer to a normal distribution than those of the untransformed values). The SNP values 11 (homozygous for the major allele), 12 (heterozygous) and 22 (homozygous for the minor allele) were transformed into numerical "dosage scale", corresponding to 0, x and 1, respectively, where $0<x<1$. We report here mainly our results from inspecting the original Lipogen 2a which we studied in detail, but

these findings are also representative for the updated datasets Lipogen 2a and Lipogen 3. Specifically, we analyzed the Lipogen 2a data as detailed below:

We first studied the distributions of the drug responses (i.e. four lipid levels) stratified according to the patients' characteristics (sex, fasting state, age, BMI and others). We computed Pearson correlation between drug responses and patients' features. We subdivided drug responses into groups according to sex, age, fasting state variables and quantified the proportion of measurements falling in each categories.

Within the Lipogen 2a study consisting of roughly ¾ of men and ¼ of women, we observed a sex effect in the TG response. Indeed the distribution of the TG levels corresponding to the male individuals was shifted towards higher TG values in comparison with the TG distribution of females. This difference between the two distributions has a very high statistical significance (*t*-test *p*-value=1E-40). TG values also appeared to be correlated with age (correlation coefficient r = 0.19), indicating the tendency for higher TG values in older patients (Fig. 5). TG was also correlated with BMI (correlation coefficient r = 0.22). Furthermore, we investigated the impact of the fasting state on the TG levels. Only about 20% of the patients had been fasting when their blood samples were taken and we found a statistical difference between the respective lipid levels (t –test: p = 0.028).

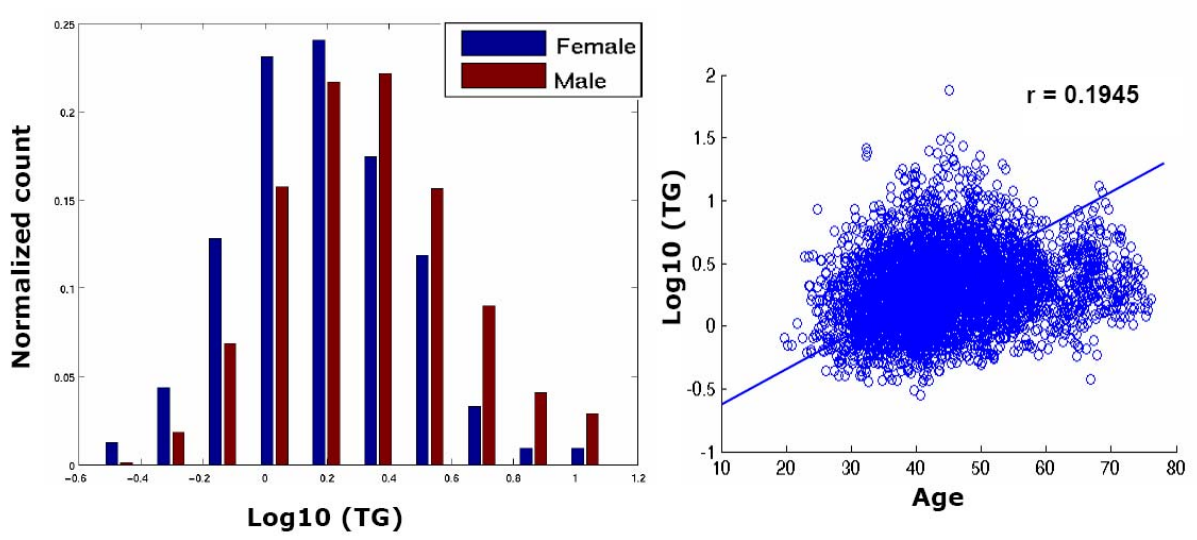


Figure 5: Log_{10} triglycerides distributions (separated by gender status) and correlation between Log_{10} triglycerides and age.

On the left plot, the blue histogram corresponds to the female $\log_{10}$(TG) and the red histogram to the male $\log_{10}$(TG). On the right plot, a scatter plot of the $\log_{10}$(TG) as function of age is represented. The blue line shows the slope of the linear regression.

For the genotypic dataset (20 SNPs genotyped in 438 patients), we confirmed that all SNPs were in Hardy-Weinberg equilibrium, with one exception (abca1-2962g) which turned out to be due to the mixture of sub-populations having different allelic frequencies for this particular SNP (the Wahlund effect [55]). We then investigated the significance of the SNP correlations (use of controls based on the reshuffling of the SNP information) (Fig. 6).

We analyzed the significance of all pairwise SNP correlations. The aim was to reveal possible groups of highly and significantly correlated SNPs. Such correlations can be due either to linkage disequilibrium (in particular for SNPs in the same gene) or due to population stratification.

To this end, we calculated all pairwise Pearson correlations between the 20 SNPs across all patients. We then performed a significance analysis by reshuffling the genotypes for each SNP 10,000 times and re-computed the corresponding correlation coefficients. We chose a cut-off of significance at 5% (requiring that less than 500 of the correlations after data reshuffling had to be higher than the observed one). We observed that SNPs in the same genes are usually significantly correlated (for example the SNPs on *APOC3*, *APOE* and *ABCA1*). However, we also found a small number of significant correlations between SNPs on different genes (mtp-493t and adrb2-16r; apoa5-64g and apoc3-455c), which are probably due to population stratification (Fig. 6).

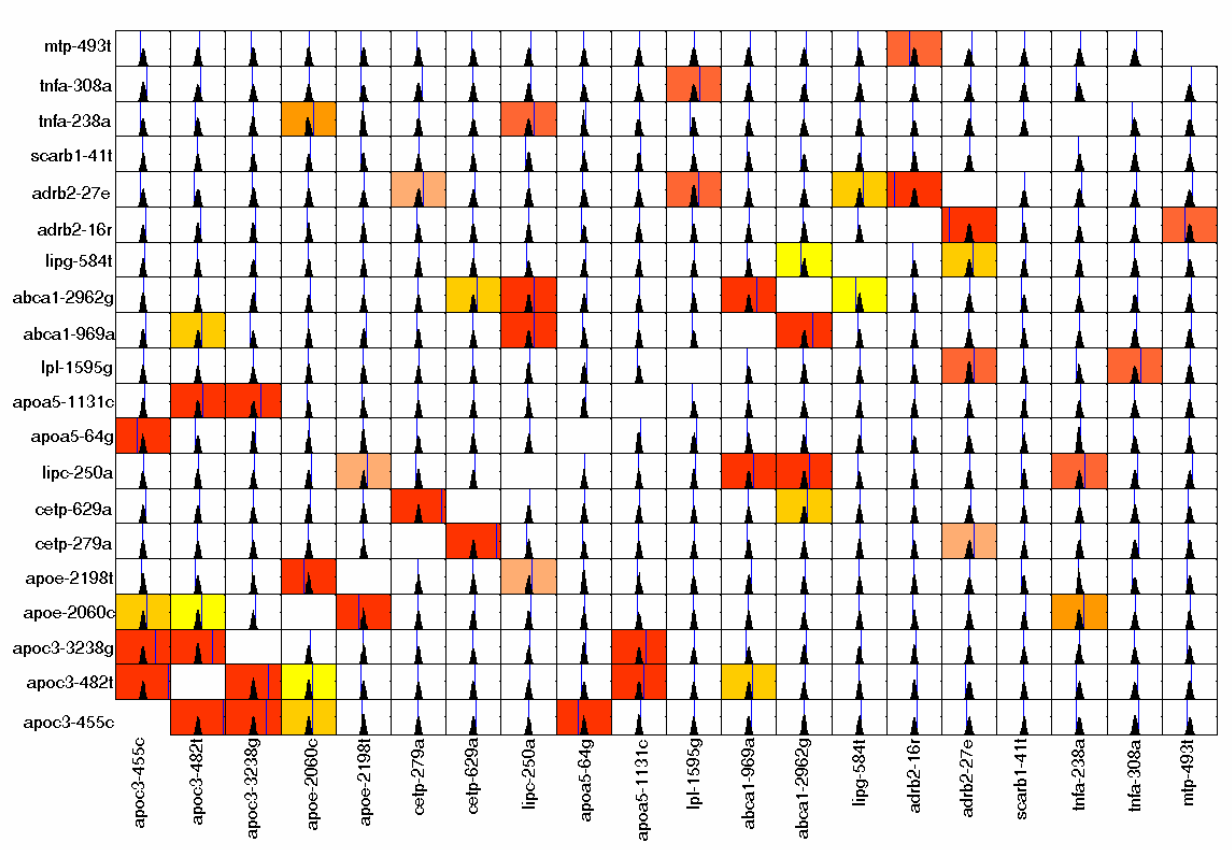


Figure 6: Significance of the SNPs correlations.

The row and column labels refer to the SNP name. Each square shows a distribution of 10 000 Pearson correlation coefficients (after data reshuffling). The blue line indicates the observed value of the correlation coefficient. The colors indicate the level of significance of the correlations ranging from yellow ($p<0.05$) to red ($p<0.0001$).

For the therapeutic dataset (18 drugs administered in various combinations for 438 patients), we studied both the treatment frequency for each drug and correlations between drugs. Specifically, we calculated all 324 pairwise Pearson correlations over all treatments. These correlations reflect the global "co-administration" indicating which pairs of drugs have been given simultaneously and which were most of the time not administered together. Our results are shown in Fig. 7 and are consistent with the practice of using combination treatment. By definition the drug autocorrelations (along the diagonal) are equal to 1, whereas the remaining pairwise correlation values range from -0.58 to 0.58, showing anticorrelation of some pairs (DDI-3TC and D4T-AZT) and relatively high correlations of some others (3TC-AZT and SQV-RTV).

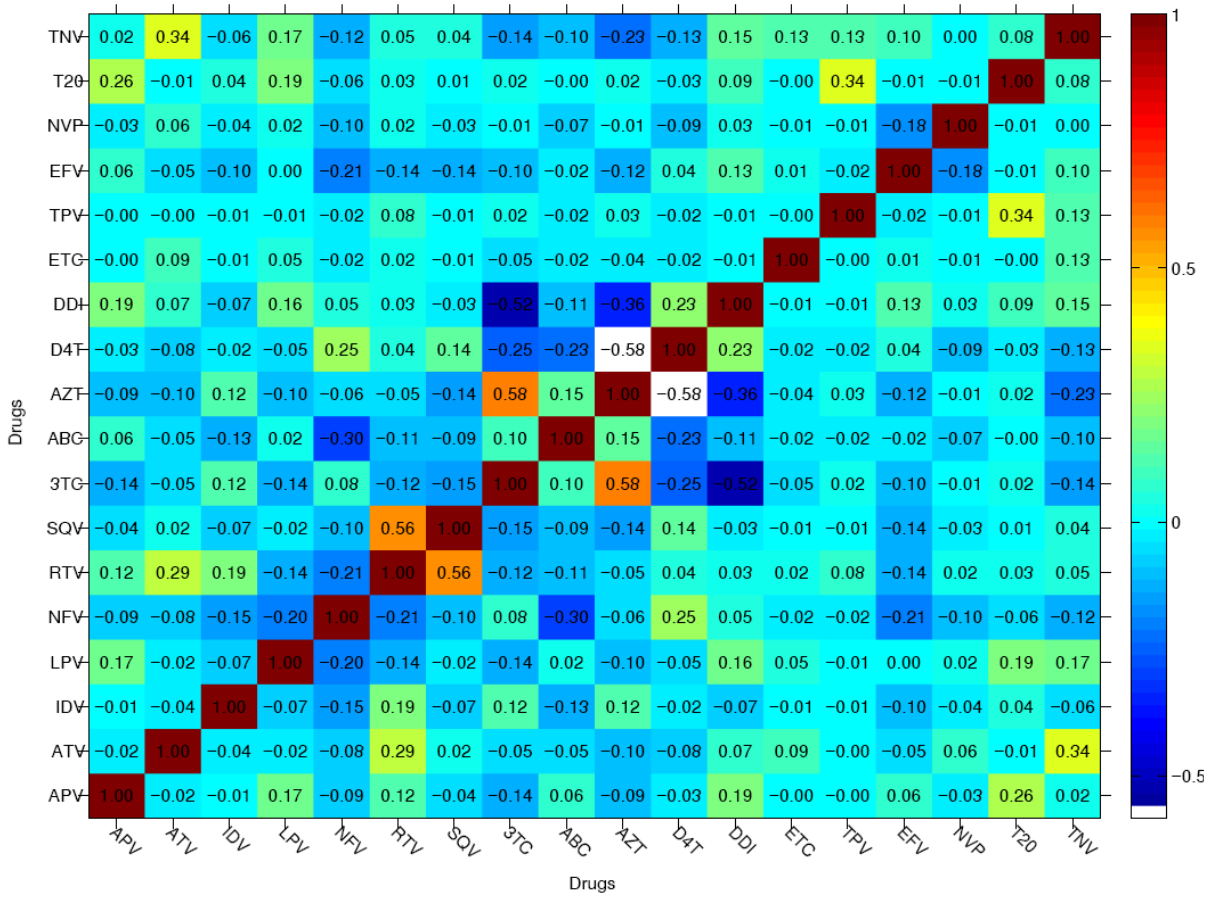


Figure 7: Pairwise drug correlations

The row and column labels refer to the drug name. Each square shows Pearson correlation coefficient. The colors indicate the level of the correlations as indicated in the colorbar on the right.

2.6.2.2 Differential response analysis

To investigate the possible effects of the feature parameters on the lipid levels from Lipogen 2a, we initially investigated the dependency of each lipid response with respect to one feature and then with respect to two features.

Response with respect to single feature

We first focused on the lipid responses as a function of one single feature, either a drug, a SNP or another characteristic of the patients. Discrete features, like the medication (drug taken or not taken), SNP genotypes or gender (male or female) divide the measurements into subgroups. We then tested the hypothesis that these subgroups have distinct underlying response distributions by performing (as an example) a *t*-test of the responses as a function of usage of a single drug (irrespective of the remaining drugs and other features).

Response with respect to two features

Next, the response profiles were investigated as a function of two features. Already for two binary features, the data can be subdivided into four subgroups. Generally, we search for "interactions" between features. For example, genetic epistasis [56] (the phenomenon where the effects of one gene are modified by one or several other genes) could imply an elevated (or decreased) response only if both SNPs have a particular allele. Similarly, certain adverse drug responses might occur only for a particular combination of drugs. Finally, some drugs may give rise to a response only in patients with a certain genotype (which is indeed the central point of pharmacogenetics as we outlined in section 2.3).

In general, epistasis is used to denote the departure from 'independence' of the effects of different genetic loci. Statistical models can be used for detection of epistasis in humans.

Epistasis in the Fisher sense is closer to the usual concept of statistical interaction: departure from a specific linear model describing the relationship between predictive factors (here assumed to be alleles at different genetic loci). Statistical tests of interaction are limited to testing specific hypotheses concerning precisely defined quantities. For statistical testing, we can only focus on mathematical models of epistasis rather than those encoded by a rather abstract and vague notion of 'independence' or 'masking' of unspecified 'effects'. Therefore, statistical interaction does not necessarily imply interaction on the biological or mechanistic level. We should like to perform a statistical test and interpret the outcome biologically, but this is in general not permissible. The degree to which statistical modeling can elucidate the underlying biological mechanisms is likely to be limited, and may require prior knowledge of the underlying aetiology [57].

In the following, we will refer to 'interaction' in the statistical sense rather than the biological one and it will be mainly in the context of SNP-drug interactions.

In order to analyze the differential response we used one feature to define a baseline. For example, the lipid measurements of patients with a certain genotype who did not take a particular drug can be used to rescale all measurements when this drug was administered. Specifically, we defined:

$$\Delta(LR)_g = LR_{g,t=1} - \text{mean}(LR_{g,t=0})$$

Here, *LR* refers to the Lipid Response (TG, HDL, NHC, Chol) for patients of genotype *g* (11, 12 or 22) at a given SNP, recorded after treatment with (t=1) or without (t=0) a certain drug.

The mean over all measurements for a certain genotype was used as a baseline level of the lipid responses since, in many cases, there was no response data with and without a given drug for the same patient. A disadvantage of this method is that the variation of the measurements without treatment is not taken into account. Given the definition of a differential response, one may then ask whether subpopulations with a certain SNP genotype respond differently to a given drug. To this end, we performed statistical tests (*t*-test and Kolmogorov-Smirnov test) that compared the distributions corresponding to different genotypes. In principle, this approach can also be applied to study SNP-SNP or drug-drug interactions by selecting first one feature to define the baseline and then the other to define the two subpopulations.

The results shown in Fig. 8 illustrate this type of analysis for the TG responses with respect to one or two features. The idea was to study the different lipid level distributions for *all* SNPs, *all* drugs, as well as their combinations. The top row and the rightmost column of Fig. 8 show the two distributions of the TG levels for each drug (taken vs. not taken) and SNP (genotype 11 vs. 12), respectively. The background colors of those plots indicate the significance for the difference between the distributions according to a *t*-test (darker plots correspond to smaller *p*-values). The remainder of the figure shows the four distributions of TG levels corresponding to a division of the data both according to all SNP-drug combinations (c.f. legend). Again, the background grey scales of the plots indicate the significance of SNP-drug interactions according to the *t*-test of the differential responses $\Delta(LR)_g$.

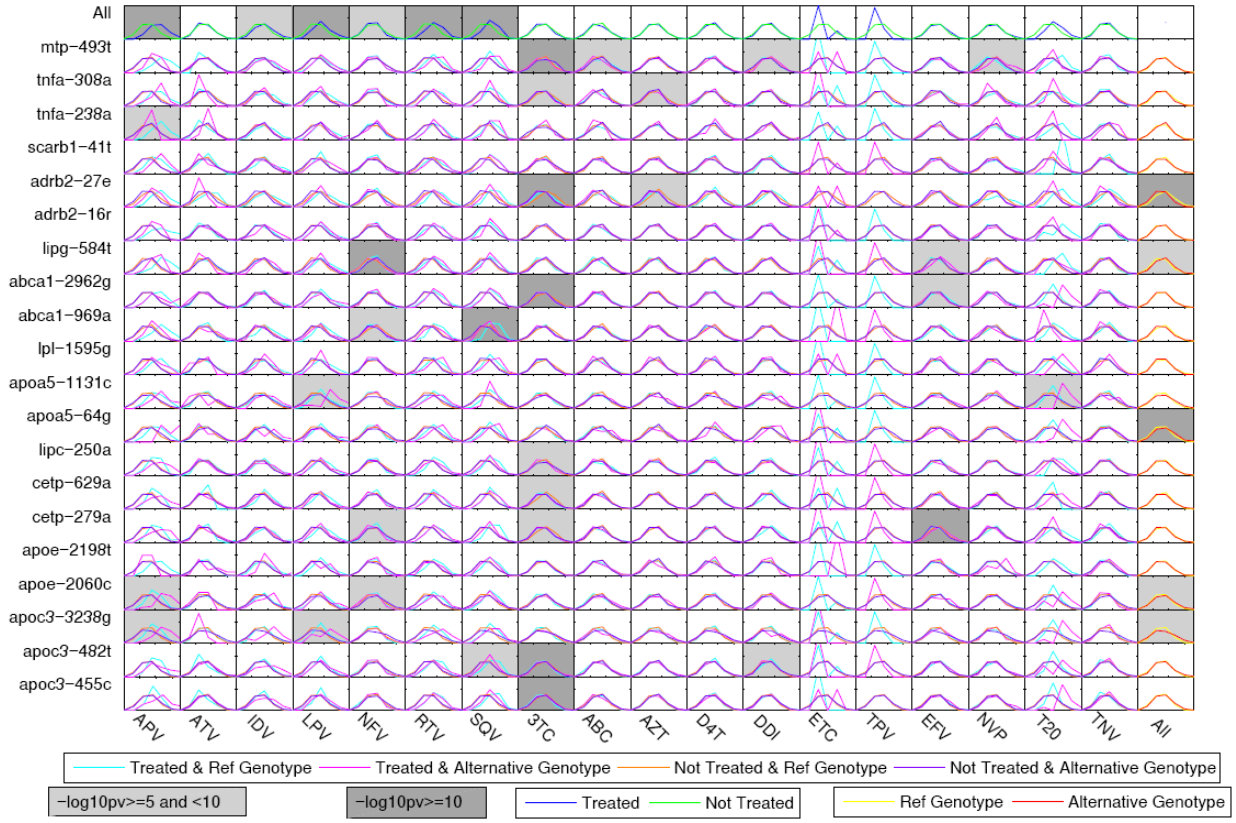


Figure 8: TG levels in different drug-SNP conditions.

The top row and the rightmost column show the two distributions of the TG levels for each drug (taken vs. not taken) and SNP (genotype 11 vs. 12), respectively. The remainder of the figure shows the four distributions of TG levels corresponding to a division of the data both according to all SNP-drug combinations. The background color of the plots indicates the significance level (legend).

We observed that some drugs, like ritonavir (RTV), saquinavir (SQV) or lopinavir (LPV), alone already have an effect on the TG, irrespective of the SNP genotypes of the patients (top row of Fig. 8). Similarly, for some SNPs (most significantly for adrb2-27e and lipg-584t) a differential TG response is detectable between the patients being homozygous for the major allele (11) and those carrying one minor allele (12) (rightmost column).

The analysis of SNP-drug interactions revealed several significant pairs (even though the TG distributions were usually not strikingly different). For instance, our analysis indicates that the carriers of one minor allele for the SNP apoa5-1131 receiving lopinavir (LPV) are more at risk for elevated TG levels than those having the 11 genotype. Conversely, the minor allele of the SNP cetp-279a appears to be protective for individuals receiving the drug nelfinavir

(NFV) (Fig. 8). We noted that this type of analysis is likely to be underpowered, either since the drugs were given only rarely (ETC, TPV and T20), or because they were given in conjunction with different combinations of others drugs, possibly confounding the lipid response.

2.6.2.3 Two-stage model selection

General purpose and requirements for linear regression

In order to overcome the limitations of our differential response analysis, we explored an alternative approach to analyze the response data using linear regression models of the following type:

$$LR_{g,t} = a + b_s\, G_s + c_d\, T_d + d_{s,d}\, G_s\, T_d + \varepsilon \qquad (1)$$

Here G_s denotes the genotype of the SNP s (usually decoded as the dosage of the minor allele, but other models were also tested as detailed below), T_d refers to the treatment with drug d (1 if the drug was given and 0 if not) and the last term allows for modeling a possible SNP-drug interaction. The goal is to estimate the different coefficients a, b_s, c_d and $d_{s,d}$ that best fit the data (in the sense that the sum of squared residuals over all data-points is minimal). We were particularly interested in SNP-drug interactions affecting the lipid response. These correspond to coefficients d whose estimated confidence interval was inconsistent with a vanishing value. The likelihood of the interaction is also reflected by a small p-value. All pairwise SNP-drug p-values can then be used to identify groups of significantly interacting SNPs and drugs.

In the context of linear regression, it is important to specify how to decode the three possible genotypes of a particular SNP. We transformed the genotypes 11 (homozygous for the major allele), 12 (heterozygous) and 22 (homozygous for the minor allele) into a numerical scale corresponding to different types of genetic models as follows: The *additive* model assigns (0 1 2) to the three genotypes implying that each copy of the minor allele is assumed to assert the same effect. In a *dominant* (0 1 1) or *recessive* (0 0 1) model the coding is such that it only matters whether at least one minor or major allele is present, respectively. The different codings were compared in terms of explained variance within simple linear regression performed for each SNP. We observed that depending on the SNP, certain codings were more

or less favorable. For instance, *additive* coding for SNPs in *CETP* gene explained more of the TG variations than other codings, whereas *dominant* coding for SNPs in *ADRB2* and *SCARB* genes explained more of the TG variations. We nevertheless decided to use subsequently only the simplest *additive* model for all SNPs because the differences were not large.

Regression analysis is particularly sensitive to missing values and the normalization of the data. Therefore, we processed the data as follows:

1. Imputation of missing genotypes and missing phenotypes: As previously mentioned, four SNPs (adrb2-27e, tnfa-238a, tnfa-308a and mtp-493t) were removed from the genotypic dataset because of their high numbers of missing values. For the remaining SNPs the mean value of the patients' genotypes (assuming an *additive* model) was computed and this value was used to substitute for the missing values. A similar strategy was applied to variables used as confounding factors when they were not recorded.
2. Normalization of all the data: Drug responses, patients' characteristics, genotypic data, treatment data and the SNP-drug interaction terms were normalized with respect to the corresponding mean and standard deviation (z-scores). The four $\log_{10}$ normalized lipid levels were also adjusted for two major confounding factors (sex and age), using multi-linear regression.

Two-stage model selection

The general aim was to explore how the lipid responses (LR) depend on the SNP genotypes (G), the treatment (T) and, importantly on interactions between a SNP and a drug (GT). To this end, we developed a two-stage model selection system using a multiple linear regression approach [22] (Fig. 9).

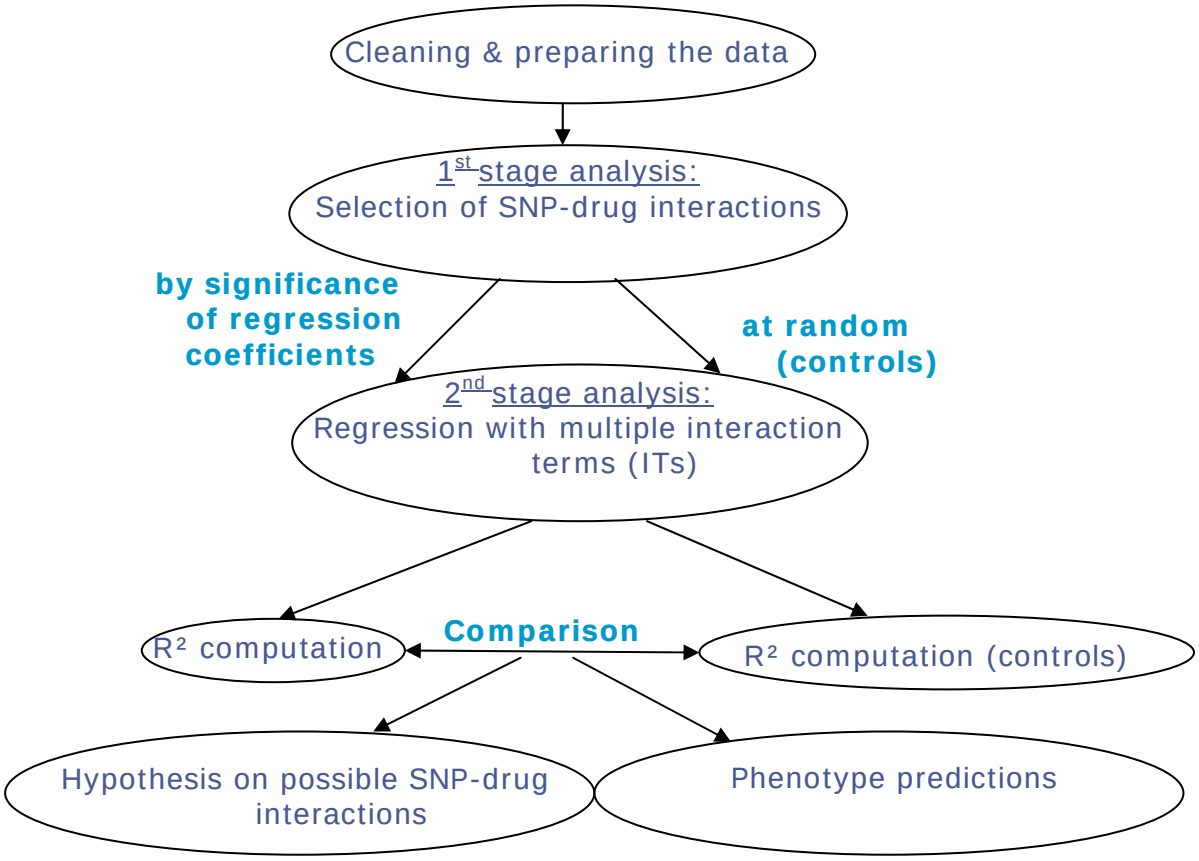


Figure 9: Overview of the two-stage analysis

The methodology used in this analysis was applied to the four lipid responses independently. Importantly, SNP-drug interactions included in the regression with multiple terms (2nd stage) were chosen either based on the significance of the pairwise interaction coefficients (from 1st stage) or at random among all possible interactions. Multiple term models were compared to control models in terms of explained variance (R^2) leading to hypotheses and response predictions regarding the selected SNP-drug interactions found at the end of the two-stage analysis.

Stage I

First, 240 independent linear regression analyses were studied to estimate the effect of only one SNP (s = 1,…, 16) and one drug (d = 1,…, 15) as well as their interaction using eq. (1). In order to assess the likelihood for such an interaction we computed a score for each $d_{s,d}$ coefficient to be non-zero. This score is defined as the distance of the estimated value of $d_{s,d}$ from zero in units of half of its confidence interval $\Delta d_{s,d.}$:

$$S_{s,d} = |d_{s,d}| / |\Delta d_{s,d}| \quad (2)$$

The scores are proportional to the absolute value of the *t*-statistics from the *t*-test testing whether the coefficients of a linear regression differ significantly from zero.

Stage II

In the second stage, we performed a multi-linear regression, using all the 16 SNPs, the 15 drugs and a subset of the most significant SNP-drug interactions (*s,d*) for which the significance score $S_{s,d}$ (computed in the first stage) is above a chosen threshold S_{min}:

$$\mathrm{LR} = \mathrm{a'} + \sum_{s=1}^{16} \mathrm{b'}_s \mathrm{G}_s + \sum_{d=1}^{15} \mathrm{c'}_d \mathrm{T}_d + \sum_{(s,d)} \mathrm{d'}_{s,d}\ \mathrm{G}_s.\mathrm{T}_d + \varepsilon \quad (3)$$

Note that this single global regression model has many more parameters (between 32 if no interaction terms are considered, up to 272 if all of them are included). In general, it is not required that the values for parameters of this model (a', b'_s, c'_d and $d'_{s,d.}$) coincide with the corresponding values estimated in the first stage.

To study how many interactions were likely to influence the response, we used the entire Lipogen 2b study (6183 measurements) to train models containing linear terms for all SNPs, and all drugs, but only certain number of bi-linear SNP-drug interaction terms. For each model, the estimated coefficients a', b', c' and d' (minimizing ε^2 in equation 3) were used to compute "fitted" lipid responses:

$$\mathrm{LR}_{\mathrm{fitted}} = \mathrm{a'} + \sum_{s=1}^{16} \mathrm{b'}_s \mathrm{G}_s + \sum_{d=1}^{15} \mathrm{c'}_d \mathrm{T}_d + \sum_{(s,d)} \mathrm{d'}_{s,d}\ \mathrm{G}_s.\mathrm{T}_d \quad (4)$$

We then checked the modeling hypothesis by comparing the fitted lipid responses to the actual responses and computing the residual values of the regression, which reflect the part of the phenotype not explained by the model.

The challenging question we had to address was how to evaluate models containing different numbers of interactions. To this end, we computed the proportion of variance in the data (including n lipid measurements) that is explained by the final model (R^2) for different selected values of S_{min}:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (LR_{i,computed} - LR_{i,data})^2}{\sum_{i=1}^{N} (LR_{i,data} - \overline{LR_{data}})^2} \tag{5}$$

where $LR_{i,computed}$ correspond to either the fitted or the predicted lipid responses.

We also computed the scores of the final interactions terms, with the estimated coefficients $d'_{s,d}$ similar as in (2):

$$S'_{s,d} = |d'_{s,d}| / |\Delta d'_{s,d}| \tag{6}$$

Comparing the two-stage model against control models

To measure the performance of our two-stage method we compared each global model with a number of pre-selected interactions with a collection of 1000 control models: These control models also included all linear terms (for SNPs and drugs), as well as a *randomly* selected set of interaction terms containing the same number of terms as the two-stage model.

R^2 values were then computed for each of the 1000 control models, and their distribution was compared with the R^2 value of the two-stage model. This procedure was repeated for different numbers of interaction terms corresponding to different chosen values of S_{min}. In order to find the "best" model we computed the difference in R^2 between each model for a given number of interactions and the mean of the corresponding control models (in units of their standard deviation):

$$R^2\text{distance} = \frac{R^2{}_{2-\text{stage model}} - \text{mean}(R^2{}_{\text{random model}})}{\sqrt{\text{var}(R^2{}_{\text{random model}})}} \tag{7}$$

This R^2 distance can provide the number of interactions where the difference between the R^2 of the model and the controls is the highest:

The two-stage model selection procedure was developed in order to identify the most essential features in the dataset, while excluding non-significant or redundant terms. It allowed for reducing the complexity of the data, especially in the large SNP-drug interaction space. Figures 10 to 13 show our results from running the two-stage analysis.

Increasing the number N of interaction terms in the model by steps of 20, we observed that the R^2 increases from around 0.15 (N=0) to 0.34 (N=240). In order to assess the significance of the increase in R^2 of the two-stage model selection method, we asked for which N the two-stage model gives R^2 values that are significantly higher than that of corresponding control models with the same N. We performed 1000 random tests. The R^2 mean of the ensemble of controls at each threshold or number of interaction terms was always below (or equal if no or all terms were included) the R^2 value computed in the two-stage model selection (Fig. 10). The difference in units of standard deviation from the random tests was relatively large for models containing between 20 and 160 interaction terms. For models including more interaction terms, the R^2 line passed below the upper "whisker" of the boxplot (which represents the largest observed value that is less than 1.5 of the interquartile range (the distance between the 25^{th} percentile and the 75^{th} percentile)). Thus, only the two-stage models that included less than three quarters of all possible interactions fitted the data significantly better than models that used the same number of randomly picked interactions (Fig. 10).

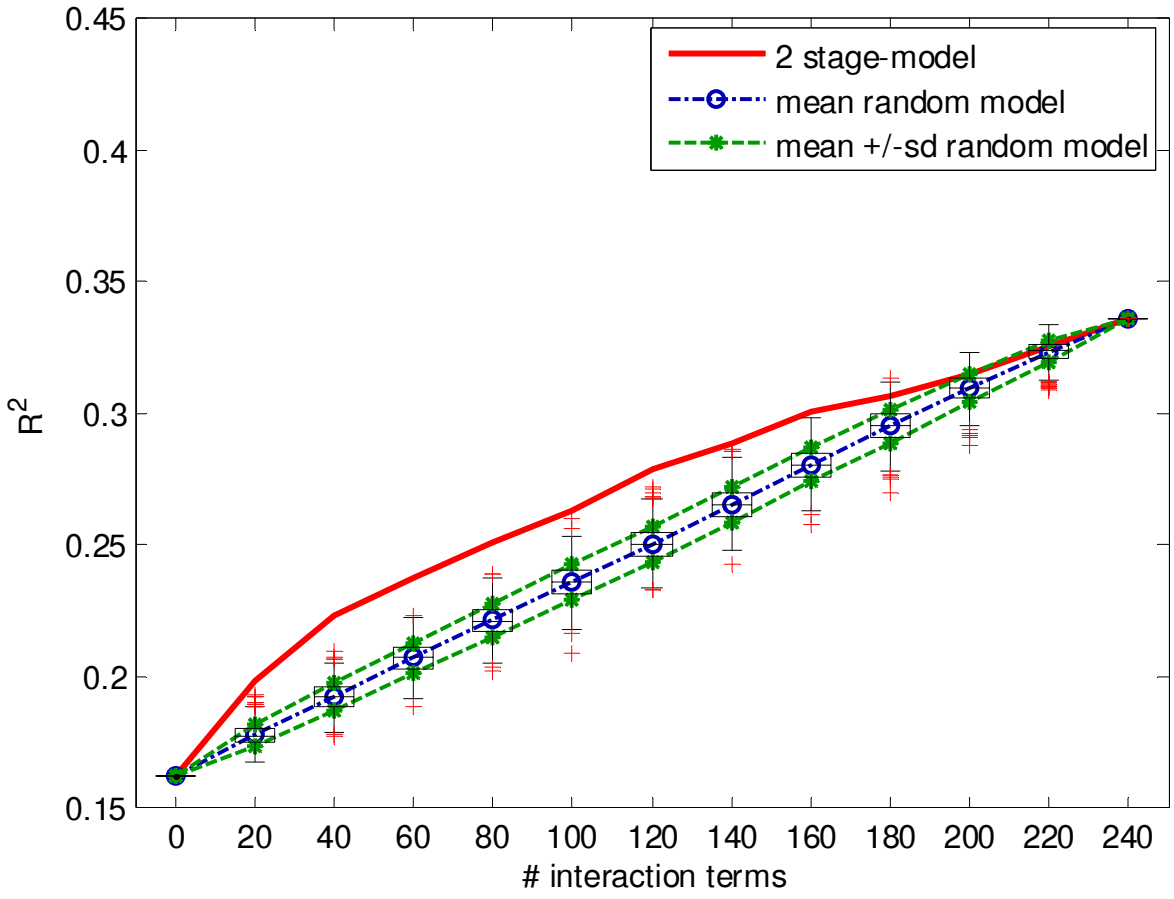


Figure 10: R^2 as a function of the number of interaction terms.

The red line shows R^2 as a function of the number of interaction terms using the two-stage model selection. The blue line corresponds to the mean R^2 of the 1000 control models with the indicated number of randomly chosen interactions. The green line shows the mean +/- the standard deviation of R^2 within the random model. The box plots show the median value as a line and the first (25th percentile) and third quartiles (75th percentile) of the distribution as the lower and upper parts of the box. The crosses represent outlying R^2 values.

To study in more detail which of the two-stage models (or essentially how many SNP-drug interactions) achieved the most improvement with respect to the controls, we computed the difference between the R^2 value for each model and the corresponding mean R^2 value, in units of standard deviation from the 1000 control models with randomly selected interactions (Fig. 11).

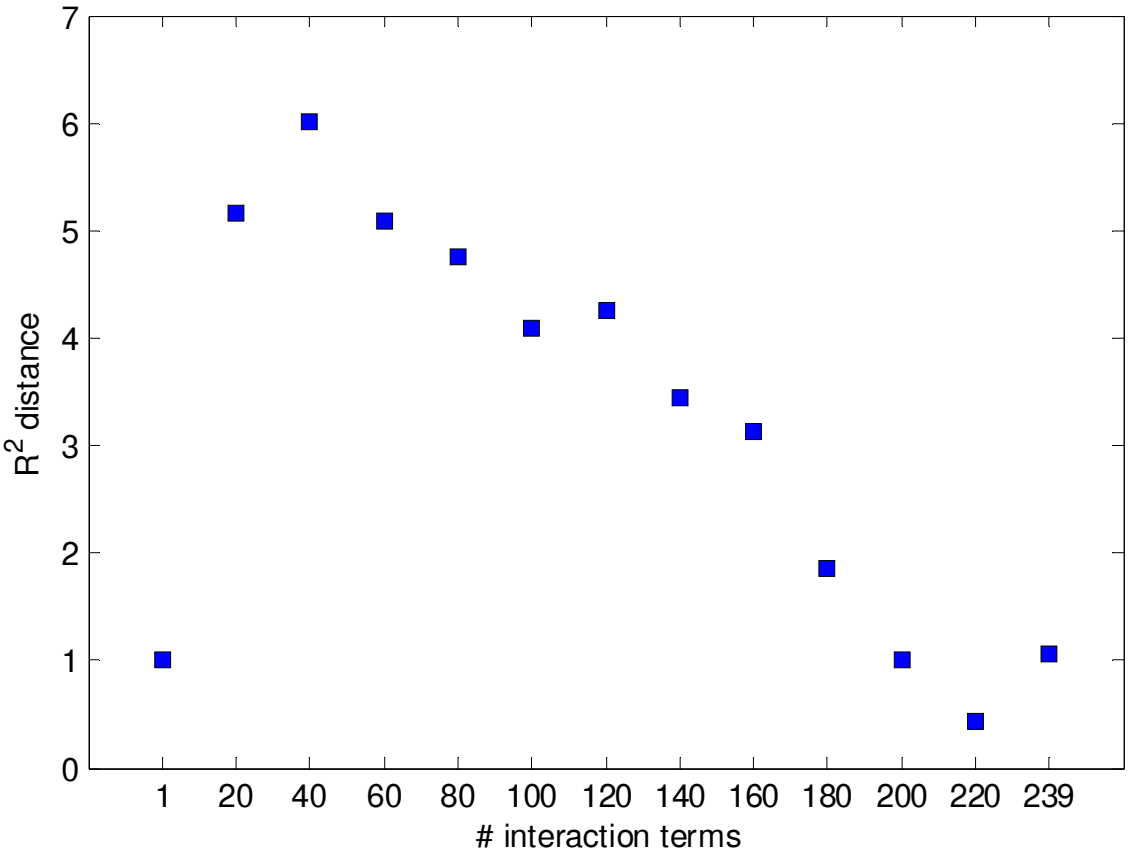


Figure 11: R^2 distance as a function of the number of interaction terms.

Each square represents the R^2 distance at each threshold S_{min} (from a model containing one interaction to a model having 239 interactions).

The R^2 distance is not defined for models with no or all interaction terms because all control models in this case are equal and thus have a vanishing standard deviation. Trivially when including or excluding only one interaction term the R^2 distance is small, because the control models are still very similar to the reference model (and to each other). We observed that the R^2 distance was also very small when using a large number of interaction terms (200 or 220), while it increased rapidly as function of interactions and peaked at 40 with a value of ~6. Adding 10 interactions also led to a R^2 distance peak (data not shown), corresponding to another local optima. When adding more than 40 terms, the R^2 distance decreased in a roughly linear fashion between 60 and 220 interactions. We concluded that allowing only ~1/6 of the possible SNP-drugs interactions in our two-stage model gave the most significant improvement with respect to the randomized controls. The R^2 value of the optimal model together with a histogram of those of the randomized controls is shown in Fig. 12.

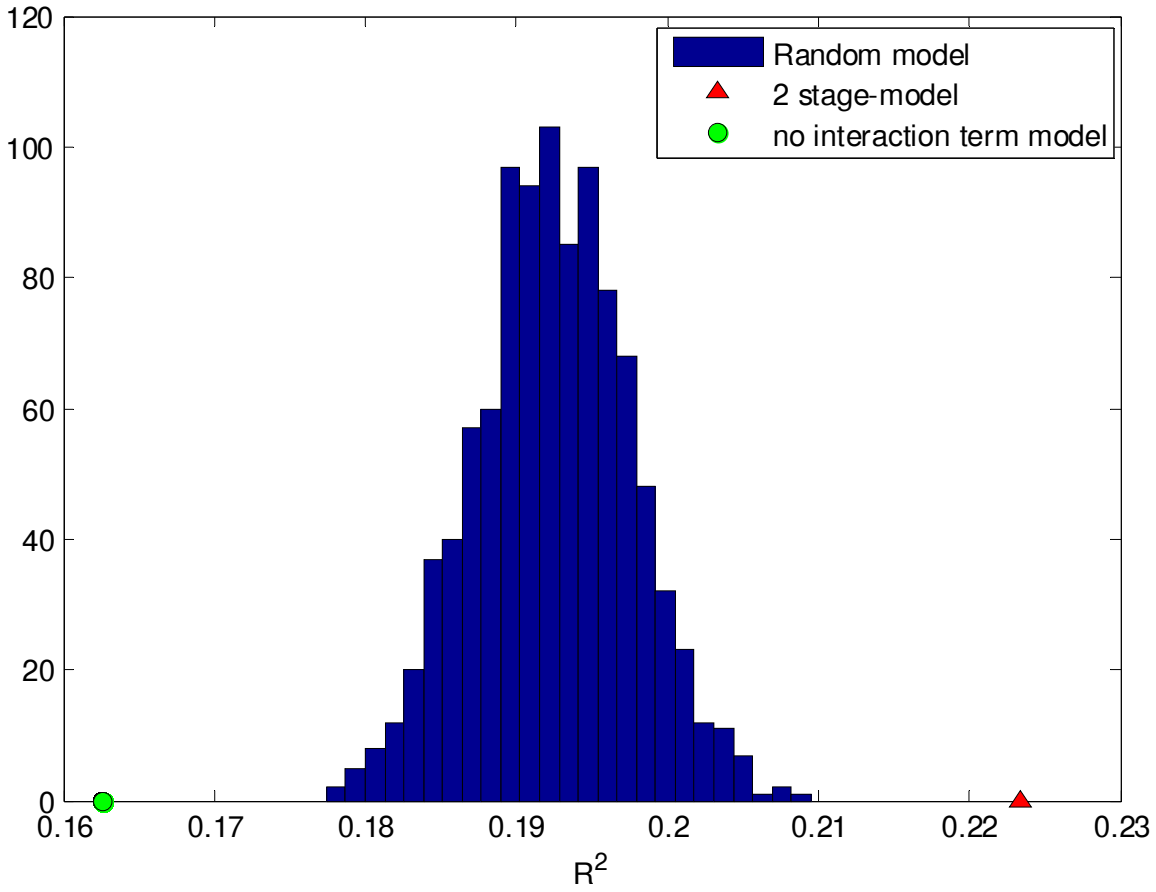


Figure 12: Distribution of the R^2 values from 1000 control models with 40 randomly picked interactions
The green circle represents the R^2 value (~0.16) when *no* interaction terms are included in the model. The red triangle shows the R^2 value (~0.22) when a 2-stage model selection is used. The *p*-value for obtaining this result by chance is below 1/1000 (<0.001).

To further assess the fit of the linear model, we compared the fitted triglyceride responses versus the actual responses (Fig. 13). The data were represented as a cloud of points, with each point corresponding to one measurement. We observed that the dynamic range of the fitted values was smaller than that of the actual values. Clearly while our model tends to capture the trend of the triglycerides, it cannot fully reproduce the measured variation. This is not surprising given that our linear model by definition cannot account for non-linear effects. Moreover, it is likely to miss out on many important explanatory variables (e.g. SNPs or environmental variables that were not measured but affect the responses). We also checked the distribution of the residuals as a function of the fitted triglyceride values, where we observed that they were approximately normally distributed, centered around zero and did not show any particular trend (data not shown).

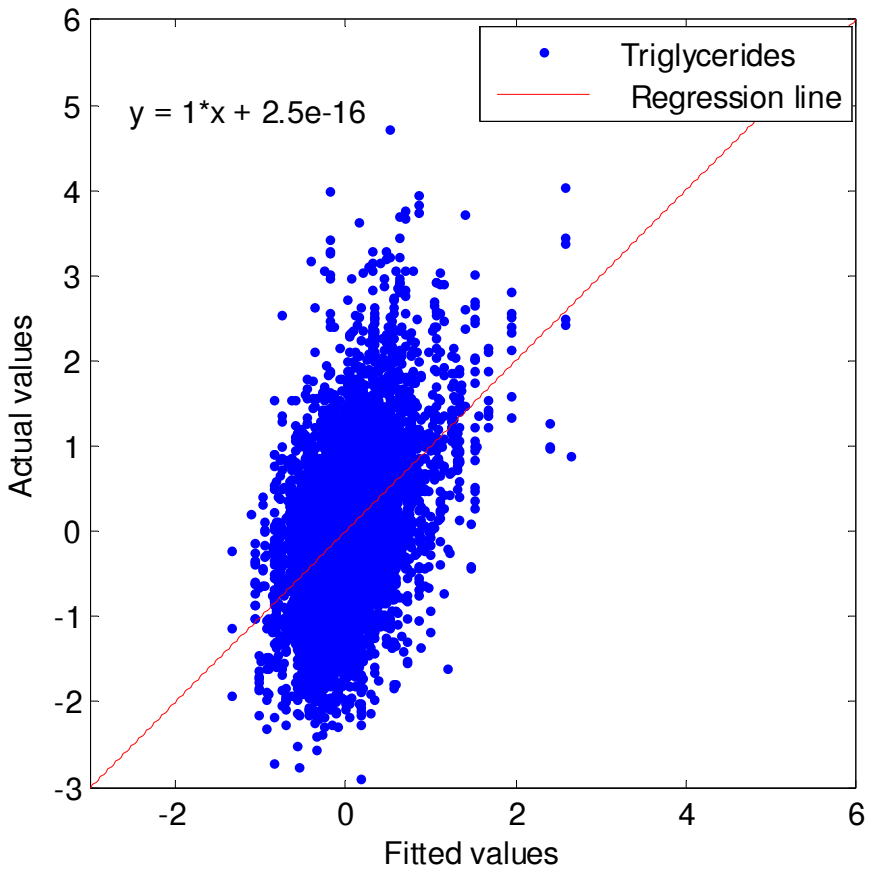


Figure 13: Actual TG values as a function of the fitted TG values

A comparison of the measured triglyceride levels with our model containing 40 interaction terms. Each blue point within the cloud of data points corresponds to a single data point. The red line has a slope of one and an intercept of zero.

A list of the ten most significant interactions (based on the scores $s_{d'}$ of our two-stage model with 40 interactions) is presented in Table 5. For each of these interactions we also showed the corresponding rank obtained in the first stage and the associated score s_d. We note that the rank obtained in the first stage was not very predictive of the rank in the second stage. A plot of the ranks of the scores at the second stage versus the ones at the first stage did not show any particular correlation (r=0.02) (data not shown). For example, the most significant interaction involving the SNP named abca1-969a and the drug SQV was ranked only at position 12 in the first stage. The exception is the interaction APV/apoc3-3238g which was among the top 10 in both the first and second stages.

Interaction names	Rank 2nd stage	Score 2nd stage	Rank 1st stage	Score 1st stage
SQV/abca1-969a	1	-3.48	12	-2.36
NVP/apoa5-64g	2	-2.84	36	-1.79
ABC/apoa5-64g	3	2.64	22	2.01
ATV/apoe-2060c	4	-2.30	15	-2.23
ABC/cetp-629a	5	2.27	24	1.96
ABC/adrb2-16r	6	2.22	23	1.97
NVP/apoc3-455c	7	-2.11	38	-1.75
D4T/abca1-969a	8	2.08	11	2.37
EFV/cetp-279a	9	1.98	29	1.87
APV/apoc3-3238g	10	1.95	10	2.41

Table 5: The ten most significant SNP-drug interaction terms.

The name of the SNP-drug interaction is shown in the first column. The rank of these interactions after the two-stage analysis in a model with 40 interactions is given in ascending order in the second column with the corresponding scores in the third column. The rank of the same interactions but after the first stage (where all the possible interactions were tested) is given in the fourth column and the corresponding score after the first stage is shown in the last column. The significance scores (which are positive by definition) have been multiplied by the sign of the regression coefficients to indicate the direction of the effect due to the interaction.

The signs in front of the significance scores are those of the regression coefficient associated with the corresponding interactions. We observed that the sign of the coefficients for each interaction is consistent between the first stage and the second stage, which was not necessarily expected since the two steps are independent. These coefficients can be interpreted as follows: a negative score decreases the risk of having elevated triglycerides and a positive score increases the chance of having elevated triglycerides. For example, patients who carry one or two copies of the minor allele of the SNP abca1-969 and were treated with SQV have (on average) a lower triglyceride level than patients who are homozygous for the major allele and were treated with the same drug. Thus, in this case the rare allele is "protective" against high TG under treatment with SQV. In contrast, patients carrying the minor allele of the SNP apoa5-64g and under the ABC treatment had, on average, higher triglyceride levels than patients homozygous for the common allele. In this case, the major allele can be seen as a "protective" allele against high TG under treatment with ABC.

Comparison with the other phenotypes

The two-stage model selection was applied in a systematic and independent manner to each of the three other phenotypes: NHC, HDL and total cholesterol (data not shown). For the TG and NHC analyses, about 40 significant SNP-drug interaction terms maximized the R^2 distance while for HDL and total cholesterol less (20) interaction terms appear to maximize this measure.

We observed only mild consistency for the selected SNP-drug interaction terms across the different lipid responses (see Table 6), except between cholesterol and NHC, where 17 out of 20 interactions are the same. With respect to the proportion of variance explained by the model (R^2), the three other lipids received lower R^2 values than TG, regardless of the number of selected interactions. Yet, the behavior of R^2 increase from a model without interaction to a model with many (selected) was comparable between the responses, with a slightly better improvement for the NHC and TG responses.

# common SNP-drug interactions	TG (40)	NHC (40)	HDL (20)	Chol (20)
TG (40)	40	13	4	7
NHC (40)	13	40	5	17
HDL (20)	4	5	20	4
Chol (20)	7	17	4	20

Table 6: Number of common SNP-drug interaction terms between the phenotypes.
The numbers correspond to the fraction of common SNP-drug interaction terms between two lipid responses for which the two-stage analysis has been run. This table can be seen as symmetric matrix. The diagonal is the total number of significant SNP-drug interactions retained at the end of the analysis. This number is also shown in parentheses next to each phenotype name**.**

2.6.2.4 Assessment of predictive power of the two-stage approach

So far our analyses focused on the question of how much of the lipid variability can be *explained* by linear models in terms of genotypic (SNPs) and environmental (drug treatment) variables, as well as possible interactions between them. Yet, it is important to realize that models that explain a certain fraction of the data variability do not necessarily have predictive

power. To establish the latter one has to demonstrate that a model that was established based only on a subset of the entire dataset also explains some of the variance of the remaining data. Within the context of the HIV data this corresponds to demonstrating that a model trained on the available data could accurately predict the expected lipid responses of *new* patients based on their genotypes for different possible treatments.

Cross-validation is the standard technique for assessing how the results of a statistical analysis will generalize to an independent data set. Its goal is to estimate how accurately a predictive model will perform on out-of-sample data that was not used for training. In general cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the test set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds [22, 58]. In *k*-fold cross-validation, the original sample is randomly partitioned into *k* subsamples (or folds). Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k*-1 subsamples are used as training data. The cross-validation process is then repeated *k* times, with each of the *k* subsamples used exactly once as the validation data. The *k* results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

There are several ways to evaluate the predictive power of a model once a cross-validation has produced out-of sample predictions. First, one can evaluate R^2 as defined in eq. 5 using the predicted responses as computed data (rather than the fitted values as we did when evaluating the explanatory power). We note that in this case R^2 can assume negative values indicating that the mean distance between predicted and measured values is larger than that between the measured values and their mean value. So actually always using the mean-value for predictions would have better predictive power than a model producing negative R^2. To avoid such negative values in the cross-validation context one often uses the so-called *coefficient of determination*, which is simply the square of the correlation *r* between observed and predicted values. This coefficient is positive by definition and a somewhat less conservative estimate of predictive power, since it tests how well the ranking of the

observations are preserved, rather than the actual values. In case of in-sample (fitting) evaluation r^2 is equal to R^2 as defined in eq. 5.

Both r^2 and R^2 only provide a global assessment of predictive power. These measures do not allow for dissecting whether the model performs well only for extreme values or if it tends to predict the correct trend for the bulk of the data. For such a more stratified analysis, it is customary to quantify the so-called *Receiver Operating Characteristics* (ROC). This technique requires a binary classification in which the individuals' measurements are classified into two groups. For example, in the context of our lipid-measurements each individual could be assigned to have either "high" or "low" lipid levels after a given treatment by applying some (arbitrary, yet fixed) threshold on the corresponding measurement. As a threshold, one could use the median value or some recommended clinical cut-off that is used to define dyslipidemia. Then to assess the predictive power of a given model one uses the predicted values to classify all measurements as "high" or "low". Importantly the dichotomization is done in a quasi-continuous manner by varying the classification threshold. This is equivalent to sorting all predicted values and then dividing the sorted list of values at all positions, starting from the beginning where all measurements are declared to be *negative* all the way to the end where all measurements are declared *positive*. At each division of the list, one can then compute the sensitivity and specificity which are statistical measures of the performance of a binary classification test. Sensitivity is the fraction of all measurements classified as positive (P) that were indeed *high*. This is equivalent to the true positive (TP) rate: TPR = TP/P). Conversely, specificity measures the proportion of all measurements classified as negative (N) that were indeed *low*. This is equivalent to the true negative rate TNR = TN/N). The relationship between sensitivity and specificity, as well as the performance of the classifier, is usually visualized in terms of the ROC curve by plotting the true positive rate (TPR) against the false positive rate (FPR), where FPR = 1 - TNR [59]. This curve allows for comparing the predictive power of all possible compromises between sensitivity and specificity. The area under the ROC curve (AUC) (also known as the ROC index) provides a single measure of accuracy (predictive power) for a given model. A perfect model has an AUC of one, whereas random classifiers give an AUC close to 0.5 [59, 60].

In-sample analysis

While ROC analysis is designed to evaluate *predictive* performance, it can also be used to characterize the *explanatory* power of a model. To this end we simply used the fitted (rather than predicted) values for classifying lipid measurements as positive or negative. Specifically we used the entire dataset to compare the two-stage model versus the control models in terms of ROC. For this end, after performing the two-stage model selection on the entire dataset, the fitted values were re-converted into their original scale (mmol / l). A clinical threshold of 2.26 mmol/l (200mg/dl) for TG was applied to divide the continuous actual lipid responses into two categories (negative or "low risk" and positive and "high risk"). This threshold was based on the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment Of High Blood Cholesterol in Adults recommendations [61], that are also used as guidelines for the HIV population [52, 62].

Using ROC analysis on fitted data allowed for an alternative evaluation of the significance of the selected SNP-drug interactions. We compared the sensitivity and specificity performance of the two-stage model versus the control model (see Fig. 14 for the TG responses of the 438 patients from Lipogen 2b with a fixed number of 40 interaction terms). We observed that the ROC curve from the two-stage model selection was above the ROC curve from the 10 control models, showing that for most of the FP rates, the TP rate of the two-stage model was higher. Consequently, the AUC of the two-stage model (0.728) was significantly higher than the mean controls (0.706+/-5.3E-3). Thus, our ROC/AUC analysis supports our claim that the selection of the 40 interaction terms using our two-stage approach had better *explanatory* power than a random picking of the same number of interactions.

However, we also noticed that the mean ROC curve for control models was significantly above the diagonal, indicating that the linear terms (for SNPs and drugs) already explain a fraction of the data and that this fraction also increases when adding randomly picked interactions. A possible explanation is that each control model may include some relevant interactions by chance.

Out-of-sample analysis

As we mentioned above, in order to evaluate the *predictive* power, we performed ROC analysis of models considering their out-of-sample estimates from various cross-validation schemes.

We made the following observation: While our in-sample ROC curve for the two-stage model had been consistently above those of the control curves generated from models with the same number of randomly chosen interactions we were not able to receive such a result for our out-of-sample predictions. Rather we observed that the corresponding ROC curves were significantly *below* the control curves (see Fig. 14) (AUC= 0.575 vs. AUC=0.633 +/-5.6E-3). This is a surprising result, because for instance the training sets of a leave-one-out cross-validation are almost identical to the entire dataset (only 1 out of 438 patients is removed at a time), but nevertheless this small difference seemed to produce a strong bias for improving prediction power with pre-selected interactions according to our two-stage approach. Nevertheless, it should be emphasized that we could still demonstrate significant out-of-sample prediction power using SNP-drug interactions, only that choosing them at random is at least as good, if not better, as picking interactions that appeared significant in a model that consider only a single SNP and a single drug at a time. One possibility might be that many of the selected SNP-drug pairs might contain redundant information explaining why random sampling of SNP-drug pairs performs better.

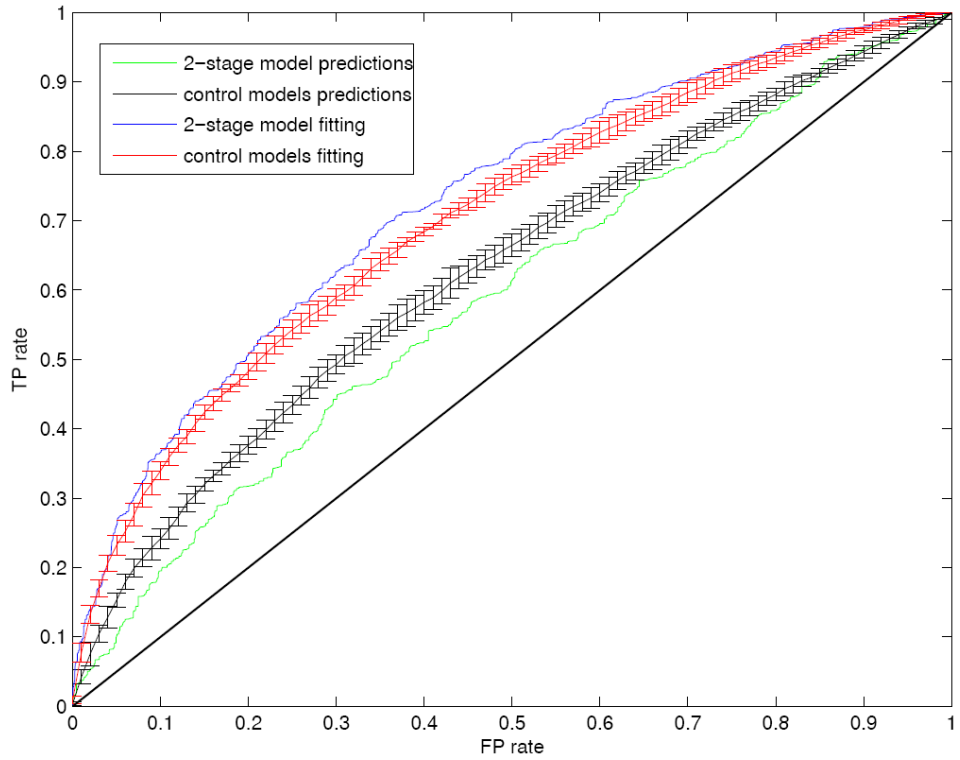


Figure 14: ROC curves illustrating performance of two-stage SNP-drug interactions model for the triglycerides response.

Shown above are the four ROC curves for fitting (blue curve) and prediction (green curve) of the the $\log_{10}$ TG, using 40 SNP-drug interaction terms. The corresponding control models are shown as mean +/- standard deviation respectively, in red and black. A 438-fold cross-validation (leave-one-patient-out) has been used for out-of-sample predictions. The clinical threshold is equal to 2.26 mmol/l.

2.6.2.5 Further refinements of the two-stage approach

In the following, we describe additional refinements to our analysis, employed to improve the predictive power of our model. However, for predicting TG values none of these modifications significantly altered our results (Fig. 14).

Correction and normalization of the response data

We applied a quantile-quantile normalization [22] to the phenotypic data before running the two-stage analysis. The transformation was applied to the lipid responses such that the transformed values had the same rank, but followed exactly a normal distribution, with zero mean and unit variance. The advantage of such a transformation is that it guarantees a normal response which leads to correct inference of the parameter estimates and often improves the performance of regression which is sensitive to outliers.

Logistic regression

Logistic regression is a useful way of describing the relationship between one or more independent variables with a binary response variable [22] (see also section 1.3). Accordingly, we therefore binarized the lipid levels prior to modeling. We then applied the same two-stage analysis on the TG measurements (split into two categories based on the classical clinical threshold) using two types of link functions (logit and probit). The AUC after cross-validation was equal to 0.57 compared to the AUC of the controls, which was equal to 0.62 +/-8.4E-3. Using a discrete TG phenotype and applying the two-stage model selection did not increase the out-of-sample prediction accuracy, compared to the corresponding control model performance.

Two-stage analysis combined with a stepwise approach

One of the limitations of the two-stage regression is that the scoring of the SNP-drug interactions at the first stage is based only on one SNP-drug pair while ignoring all other pairs, while the second stage usually includes many pairs. Some pairs may be redundant, e.g. because the respective SNPs are identical (or in high linkage disequilibrium) and the respective drugs may have been given to many patients. In this case only one pair may be attributed a significant interaction in the second stage. In principle a multi-stage method (e.g. stepwise model selection [22, 63]) may be an efficient means to identify a collection of non redundant significant interactions. Stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure (adding and removing terms from a multi-linear model based on their statistical significance in a regression; see also section 1.3). We implemented a stepwise procedure to refine the two-stage model in order to remove redundant SNP-drug interaction (while keeping also the linear terms). The final number of interactions has also been used for the control models.
This procedure was applied after the two-stage model selection to tune the final number of interactions and potentially remove the redundant ones. The procedure uses the *F*-statistic or the Akaike information criterion for selecting the best interaction terms. We computed the AIC criterion for each model based on the number of parameters and the maximized value of the likelihood function. We selected the model with the lowest AIC value. However, the resulting ROC curves and AUC values did not exhibit better performance in terms of out-of-sample predictions, compared to the control models.

Additional covariates adjustment

The lipid phenotypes may be influenced systematically by confounding factors such as gender, age, fasting state, diabetes status, smoking status, use of lipid lowering agents (LLA). In order to study the impact of these factors, we performed an analysis of all lipid responses on the whole population, as well as on sub-populations stratified by these factors.

We studied only the male population (which represents around 80% of the data) in order to see if gender has any effect on our result.
The triglycerides showed a higher predictive value (AUC = 0.6112) for out-of-sample predictions (but still below controls, AUC =0.6251+-8.6E-3) when only investigating the male population whereas the other lipid responses were not as sensitive as TG to gender difference. Similarly, we explored the impact of age, fasting state, diabetes status, and smoking status in our two-stage analysis. We found that none of these covariates helped to improve the predictive power of the cross-validation procedure for TG levels. As an example, the cross-validation AUC for non-LLA treated people was equal to 0.575 whereas the control models for the same patients gave an AUC of 0.638 +/- 8.2E-3.

To adjust the data for effects induced by LLA (which might have contributed to misclassifications), we needed to estimate what the lipid responses of patients would have been under LLA if they had been without this treatment. A reasonable assumption is this "true" lipid level has been higher than the observed one, and we used a simple method [64] that imputed "true" lipid level as the mean of all lipid values measured in the non treated population that are above a threshold. This correction was applied to roughly 10% of the measurements and the two-stage analysis was re-run on the LLA corrected phenotype.

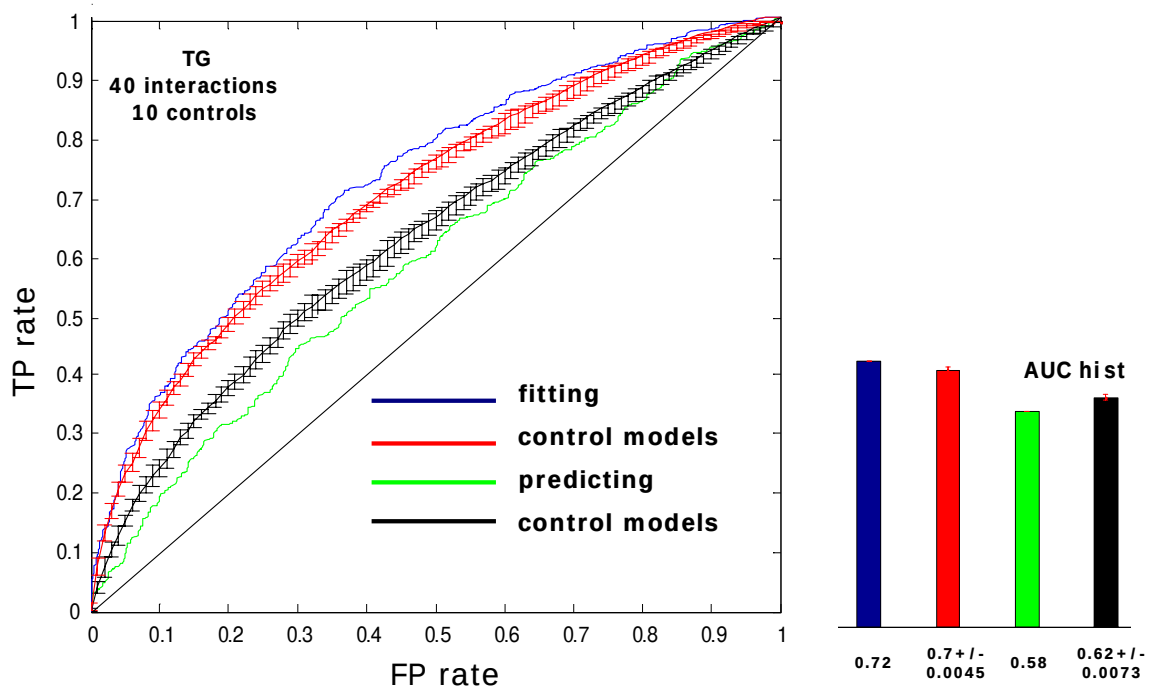


Figure 15: Predictive performance of two-stage SNP-drug interactions model for the triglycerides response.

The $\log_{10}$ TG have been corrected for sex, age, LLA variables and quantile-quantile normalized, using 40 SNP-drug interactions. A 438-fold cross-validation (leave-one-patient-out) has been used for out-of-sample predictions. The clinical threshold is equal to 2.26 mmol/l. The four ROC curves are shown on the left side and the corresponding AUC values are displayed on the right side.

Figure 15 shows the ROC analysis (438-fold cross-validation) for the two-stage model of quantile-quantile normalized TG after correction for sex, age and LLA. As previously mentioned, the ROC curve for fitted values was, for all false positive rate (FPR) values, significantly above the control curve generated from models with 40 randomly chosen interactions (AUC=0.72 vs. AUC=0.7+/-4.5E-3). In contrast, out-of-sample predictions generated a ROC curve that was significantly below the control curve (AUC=0.58 vs. AUC=0.62+/-7.3E-3).

Two-stage analysis for SNP-SNP interactions

Our approach can also be applied for studying other types of interactions, like SNP-SNP or drug-drug interactions. We tested this on the triglyceride phenotypes corrected for sex, age, LLA, which was subsequently quantile-quantile normalized. For the SNP interactions, the

second stage regression contained all 16 SNPs and a subset of all 136 non-redundant SNP-SNP interactions.

Using cross-validation, we found increased predictive power of the two-stage model over its random controls for most values of the FPR when using 20 SNP-SNP interactions (AUC=0.56 vs. AUC=0.52 +/-1.4E-2) (Fig. 16).

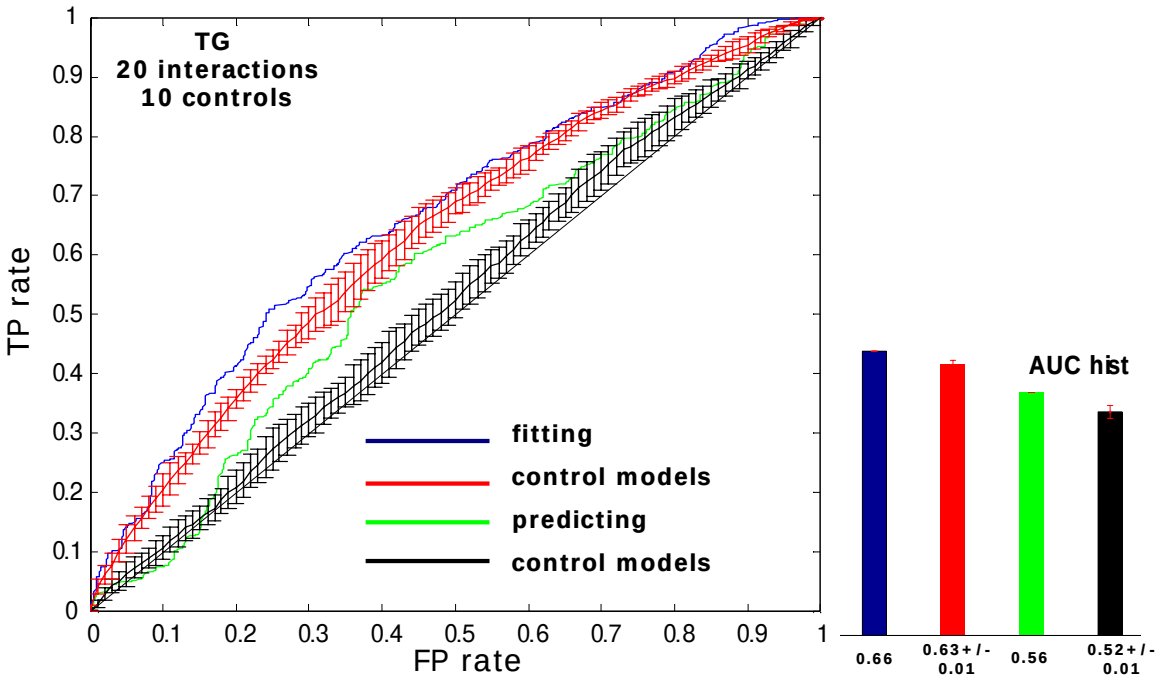


Figure 16: Predictive performance of two-stage SNP-SNP interactions model for the triglycerides response.

The $\log_{10}$ TG have been corrected for sex, age, LLA variables and quantile-quantile normalized, using 20 SNP-SNP interactions. A 438-fold cross-validation has been used for out-of-sample predictions. The clinical threshold is equal to 2.26 mmol/l. The four ROC curves are shown on the left side and the corresponding AUC value is shown on the right side.

While this is an encouraging result, we noted that the total explained variance is smaller in this case than for models that include linear terms for drugs or SNP-drug interactions. This indicates that the effect of the treatment dominates the lipid response, making it difficult to extract the potential effects of the SNPs and of the SNP-SNP interactions from this dataset.

Two-stage analysis for drug-drug interactions

For the drug interactions, the second stage regression contained all 15 linear drug terms and a subset of interaction terms among all 120 bi-linear drug-drug interaction terms. We tried interaction term subsets of different sizes using only the most significant terms from the first stage. The two-stage model suffered from rank deficiency at the level of the matrix of the explanatory variables due to linear combination of interactions terms (only 0 and 1 as possible values). The prediction results were not stable and did not show any improvement of the in or out-of-sample predictions over the control model predictions.

Two-stage analysis for gene-drug interactions

The high correlation between some of the SNPs (usually within the same gene) prompted us to average out these SNPs and re-run the two-stage analysis using only 10 genes, 15 drugs to search for relevant gene-drug interactions (among 150). This refinement did not help in improving predictive power of the two-stage model over the control models.

2.6.2.6 The two-stage model selection versus a forward model selection

The data analyses of this section were performed on the latest SHCS dataset, Lipogen 3. For a more comprehensive analysis of the predictive power of different models (detailed below), we focused on a single response to antiretroviral treatment. Specially, we focused on the $\log_{10}$ transformed triglycerides, adjusted for sex, age, LLA and then quantile-quantile normalized. We compared the following procedures which aimed at selecting a subset of the 384 (24 SNPs*16 drugs) possible interactions:

- *Two–stage regression analysis*: This procedure was described in detail before (see also section 2.6.2.3); each interaction term was first evaluated independently and then the N terms with the highest scores were included together in the full model (for different choices of N).
- *Forward stepwise procedure*: Here interaction terms are added one at a time in a greedy manner. Specifically, we used a systematic method for adding terms from a multi-linear model based on their regression coefficient statistical significance. The method started with an initial model and then compared the explanatory power of incrementally larger models. At each step, the p-value of an F-statistic (or adjusted R^2,

Akaike information criterion, Bayesian information criterion, or false discovery rate) was computed to test models with and without a potential term.

- *Control models*: These models (c.f. 2.6.2.3) provided the reference to evaluate the two other procedures. They include the same number of interactions as the reference models, but used a random selection. This procedure was repeated at least 10 times for each number of interactions.

To assess and compare the performance of the different models for predicting the lipid levels we used the explained variance R^2 (and other related metrics) as well as ROC analysis. R^2 and the area under the curse (AUC) were compared for different *N,* both for in-sample fitting and for out-of-sample predictions (cross-validation). For the latter, we tried different *k*-fold cross-validation settings: k=3, k=752 (leave-one-patient-out) and k=12170 (leave-one-measurement-out). In addition, we investigated the dependence of the cross-validation on whether measurements of the same patient were distributed among the different folds (*random* partitioning vs. *stratified* partitioning). The underlying rationale for comparing these two types of partitioning was to test whether having some data of a particular patient in the training set would have an impact on the predictive performance. This is important since in a realistic setting obviously no lipid measurements are available for new patients.

We first investigated the predictive power for *random partitioning* of the measurements using three-fold cross-validation. We found that a linear model containing only SNPs and the drugs but no interactions already explained a significant fraction of the observed lipid levels (R^2~15%). Using the predicted values for classification gave an AUC of 66.5% seemingly indicating significant predictive power of our model. When adding SNP-drug interactions using the step-forward procedure the R^2 further increased by ~3% to 18% (for 25 interaction terms) and the AUC also increased by ~1.5% to 68%. For the two-stage model, the improvements for R^2 values and AUC were below the ones from the forward model but the trend was similar (R^2 = 17% and AUC = 67.5%), see Fig. 17, upper part). In contrast adding the same number of randomly selected interactions only marginally increased the R^2 and AUC values. However, when trying to predict the lipid responses using a 3-fold *stratified* cross-validation (where data from a single patient are not distributed across different folds), the models showed instable behavior and actually performed worse than the control models. The R^2 and AUC curves decreased as a function of the number of interactions for both two-stage and forward models as well as for the control models (Fig. 17, lower part).

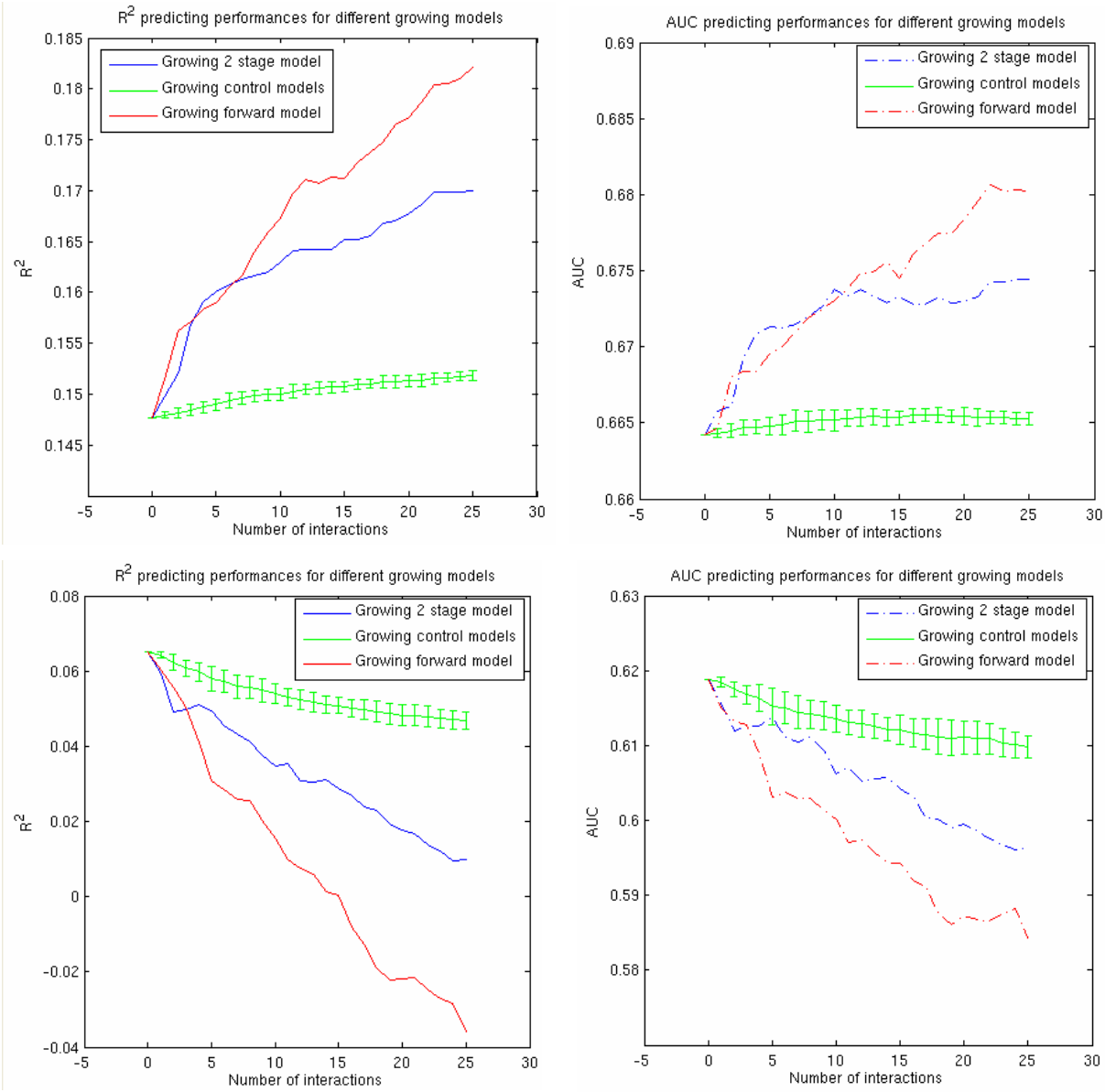


Figure 17: SNP-drug interactions R^2 and AUC of the two growing models for the triglycerides response.
The $\log_{10}$ TG were corrected for sex, age, LLA variables, using 25 SNP-drug interactions. We used a 3-fold cross-validation for out-of-sample predictions. On the upper part of the figure we represent the R^2 curve and AUC curve after *random* partitioning of the lipid responses across the folds. On the lower part of the figure we represent the R^2 curve and AUC curve after *stratified* partitioning of the lipid responses across the folds. The red color corresponds to the forward model selection, the blue to the two-stage model selection and the green to the control models.

We also observed that for 25 SNP-drug interactions, the AUC was lower than the one for the control models. The higher predictive power of the control models could be explained by the fact that in a *stratified* partitioning, each fold contains no information about the patients' measurements of the other folds, therefore, making each patient very different from the others

in terms of features present into the model. A likely explanation is that this study is underpowered to pick up interaction terms (if any) resulting in overfitting of the data. As a result, the selected interactions seem to be very specific to each patient (fold), indicating why randomly picked interaction terms showed better predictive power when the lipid measurements are distributed across the folds. The decreasing R^2 curve could be explained by the fact that cross-validation scheme can lead to negative R^2 values. In addition, for both two-stage and forward models, by definition, they were built only by inclusion (and no exclusion) of interaction terms, allowing for redundant terms and thus overfitting of the data.

2.6.2.7 In-silico data testing

We were concerned by the fact that neither our two-stage nor the forward stepwise approach for including SNP-drug interactions into predictive models had better predictive power under cross-validation than the control models containing randomly selected SNP-drug interaction terms. We thus decided to test our methods with *in-silico* data in order to rule out the possibility that there is a fundamental problem or programming error in our algorithms. Using artificial data we *know* which interaction terms are present and what their effect sizes are. Thus, we can test the functioning and the sensitivity of the various approaches. Moreover, we can compare whether the different cross-validation approaches give rise to different results as we observed for the real data.

We generated an *in-silico* dataset as follows: We first used random numbers drawn from an exponential distribution that were multiplied by random numbers drawn from a binomial distribution to define the effect sizes (and their sign) of the SNPs, drugs and SNP-drug interactions. Once such a model was established, we generated *in-silico* lipid signals using exactly the same feature variables as in 12170 measurements of the Lipogen 3 study. Finally, we added random numbers drawn from a normal distribution to simulate noise.

The data were analyzed using the two cross-validation scheme we had employed previously for the real data with either *random* partitioning or the *stratified* partitioning.

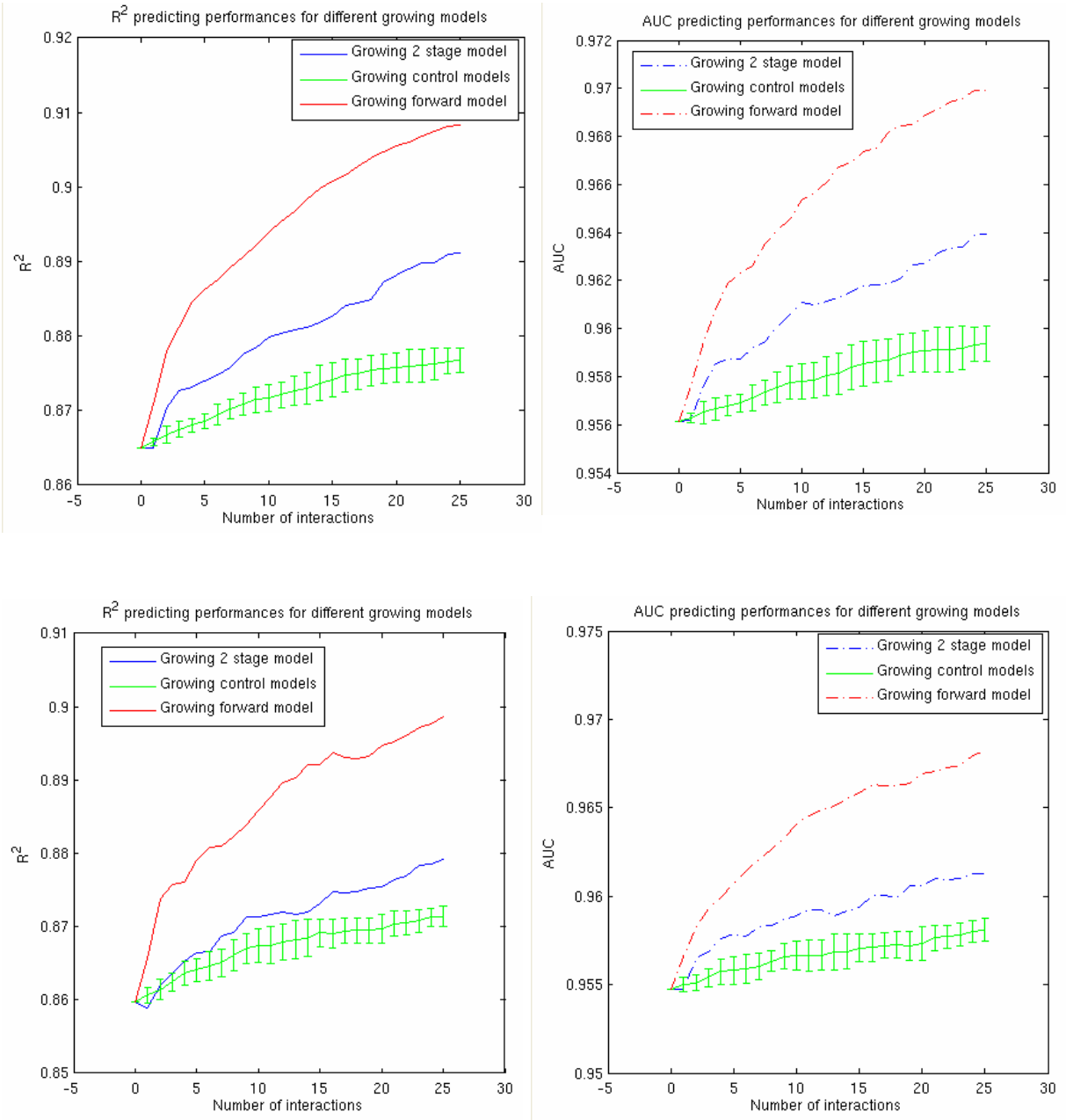


Figure 18: In-silico testing: SNP-drug interactions R^2 and AUC of the two growing models for the triglycerides response.

The $\log_{10}$ TG were corrected for sex, age, LLA variables, using 25 SNP-drug interactions. We use a 3-fold cross-validation for out-of-sample predictions. On the upper part of the figure we represent the R^2 curve and AUC curve after *random* partitioning of the lipid responses across the folds. On the lower part of the figure are represented the R^2 curve and AUC curve after *stratified* partitioning of the lipid responses across the folds. The red color corresponds to the forward model selection, the blue to the two-stage model selection and the green to the control models.

We observed that when assigning the measurements randomly to the three folds, the R^2 increased from 86.5% (no interaction) to 91% (25 interaction terms) and the AUC increased

from 95.6% (no interaction) to 97% (25 interaction terms) for the forward model. For the two-stage model, the R^2 values and AUC were lower than those from the forward model but the trend was similar. R^2 went from 86.5% to 89% and AUC from 95.6% to 96.4% (Fig. 18, upper part). Furthermore, when trying to predict the lipid responses using a 3-fold *stratified* cross-validation (keeping all measurements of a patient in a single fold) both models exhibited R^2 and AUC curves above the control models curves (Fig. 18, lower part). Similar observations were made when using different number of folds (k = 10, or k = 752 corresponding to leave-one-patient-out) or when changing the level of the noise.

We make the following conclusions:

1. The implementations of our algorithms for our two-stage and the forward stepwise approach for including SNP-drug interactions are able to identify SNP-drug interactions and give rise to predictive models that outperform the random control models for data that indeed contains signals for these interactions.
2. As expected the forward model had better performance than the two-stage method.
3. The patient specific effect is not present in the in-silico data and that could explain why the in-silico data results are not aligned with our results when performing the *stratified* cross-validation (Figs. 17-18, lower part).

2.6.2.8 Stepwise model selection of the different features and predictive assessment

To further investigate the contributions of the SNPs, the drugs and the SNP-drug interactions, we decided to analyze different models where the features (SNPs, drugs and interactions) were selected as before within the cross validation procedure (leave-one(patient)-out) using a stepwise approach (which seemed to perform better, see also section 2.6.2.6).

Stepwise model selection includes a forward selection, which involves starting with a set of variables included in an initial model, examining each variable one by one and including those which are statistically significant. Stepwise model selection also includes a backward elimination step, which given an initial set of starting variables, removes those which are not statistically significant. The stepwise approach is a combination of forward and backward steps, testing at each stage for variables to be included or excluded, based on the F-statistic [22].

The $\log_{10}$ TG responses were adjusted for various confounding factors (patient id, sex, age, fasting state, LLA, log CD4+ T cells, log viral RNA, DM2, BMI, smoking status, ethnicity, presumed mode of HIV transmission, waist). TG was regressed on each of these covariates. The new TG being used for the features selection and assessment of the model performances was defined using the residuals from the linear regression.

The following models were tested for TG predictions:
-Model with only stepwise selected SNPs terms (Fig. 19)
-Model with only stepwise selected drugs terms (Fig. 20)
-Model with stepwise SNP and drug terms (Fig. 21)
-Model with preselected SNP terms, drug terms and stepwise selected SNP-drug interaction terms (Fig. 22)
-Model with preselected SNP terms, drug terms and stepwise selected SNP-SNP interaction terms (Fig. 23)
-Model with preselected SNP terms, drug terms and stepwise selected drug-drug interaction terms (Fig. 24)

We reported the results of the ROC analysis (plot of the TPR as a function of the FPR) after fitting and cross-validating the Lipogen 3 data. Control models (as described earlier) that included the same number of interactions but chosen at random were also run to assess the model selection. Table 7 summarizes the different AUC, r^2, R^2 for those models.

We studied a model containing subsets of SNPs that were stepwise selected among the 22 possible ones. We analyzed the predictive power of this model in terms of in-sample and out-of-sample predictions (Fig. 19). The fitting curve is always above its respective controls and consequently the AUC value is higher (58.58% vs. 55.65%). The leave-one-(patient)-out cross-validation curve (below the fitting one as expected) is on average above its control line (AUC=54.8% vs. AUC=52.5%). Figure 19 shows that the SNPs selected by stepwise have predictive power and selecting SNPs performs better than picking them at random in both situations (fitting and cross-validation). In terms of R^2, altogether the SNPs explained 1.03% of the lipid level variations when cross-validating the data and 4.7% after fitting.

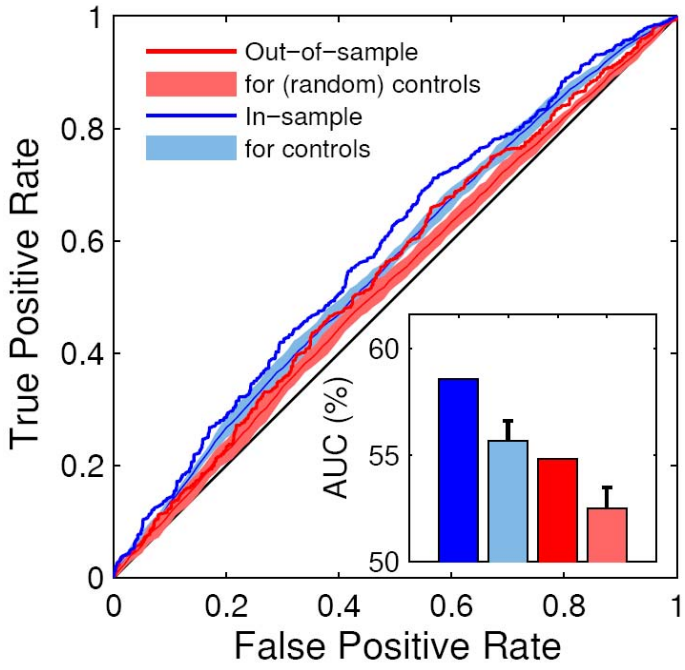


Figure 19: ROC curves and AUC values for the model containing only selected SNPs.

On the left side are represented four ROC curves and on the lower right box, the corresponding AUC values. The blue curve and bar plot refer to the in-sample predictions and AUC. The light blue ones correspond to the controls. The red curve and bar plot refer to the out-of-sample predictions and AUC. The light red ones correspond to the controls.

The same analysis has been performed using only the treatment data (Fig. 20). The fitting curve overlaps with its respective controls with similar AUC values (61 %). Figure 20 shows that the drugs selected alone in an in-sample context have a little more predictive power than the SNPs selected alone (Fig. 20). The leave-one-(patient)-out cross-validation curve shows unstable behavior, lying under its control line (AUC=56.25% vs. AUC=59.9%). Selecting the drug performs worse than picking them at random. This observation may be due to the large difference of drugs selected in each fold (patient) showing that predicting which drugs affect the TG may vary importantly from one patient to another. In terms of R^2, altogether the drugs explained 3.16% of the lipid level variations after cross-validation the data and 7.33% after fitting.

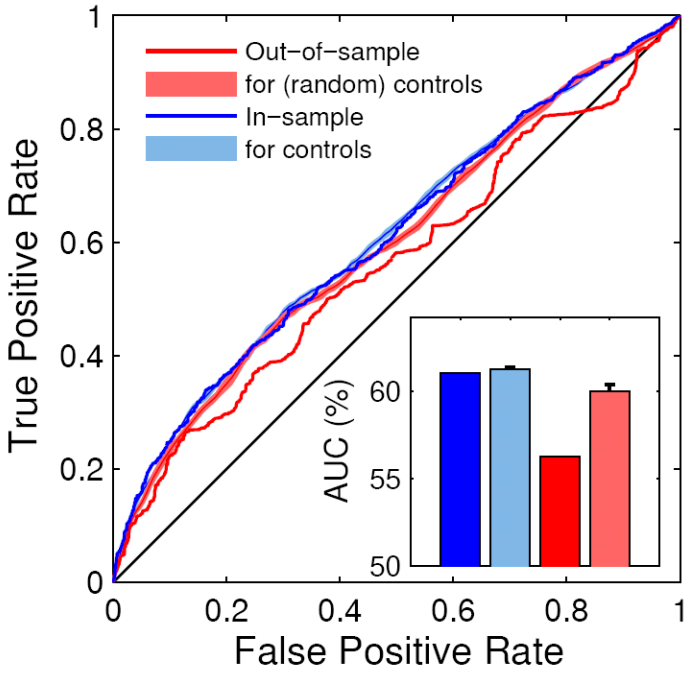


Figure 20: ROC curves and AUC values for the model containing only selected drugs.

On the left side are represented four ROC curves and on the right side box, the corresponding AUC values. The blue curve and bar plot refer to the in-sample predictions and AUC. The light blue ones correspond to the controls. The red curve and bar plot refer to the out-of-sample predictions and AUC. The light red ones correspond to the controls.

Next, we similarly investigated a model containing only SNPs and drugs selected via the stepwise approach; we analyzed the predictive power of the model in terms of in-sample and out-of-sample predictions (Fig. 21). The fitting curve is above its respective controls and consequently the AUC value is higher (66% vs. 64.27%). The leave-one-(patient)-out cross-validation curve is slightly above its control line (AUC=62.37% vs. AUC=61.85%). Figure 21 shows that the SNPs and drug selected have predictive power but the model is nearly equivalent to randomly picking the same number of SNPs and drugs from the dataset. In terms of R^2, altogether the SNPs and the drugs explained 7.63% of the variations in lipid levels after cross-validation (and 13.46% after fitting). In comparison to selecting the SNPs and the drugs from the whole dataset and fixing them before running the cross-validation, the predictive performances are better (figure not shown), revealing that the set of SNPs and drugs selected within the cross-validation process vary from one patient to another.

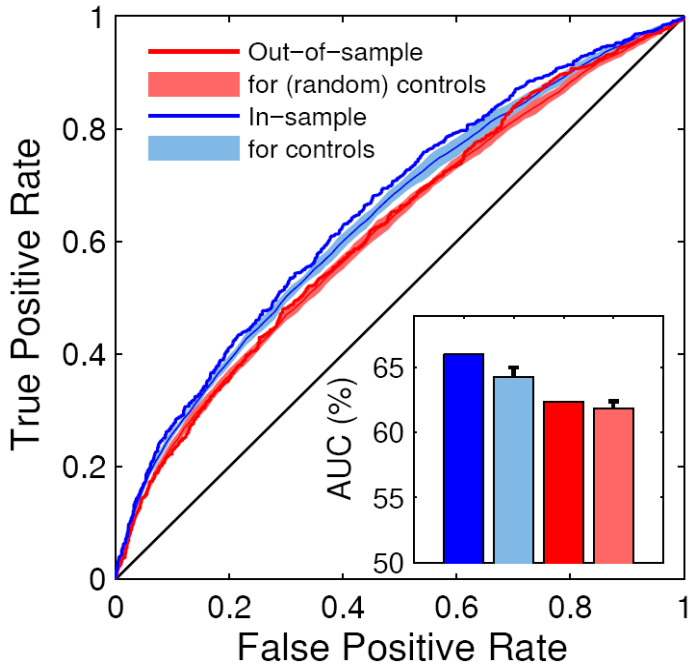


Figure 21: ROC curves and AUC values for the model containing only selected SNPs and drugs.
On the left side are represented four ROC curves and on the right side box, the corresponding AUC value. The blue curve and bar plot refer to the in-sample predictions and AUC. The light blue ones correspond to the controls. The red curve and bar plot refer to the out-of-sample predictions and AUC. The light red ones correspond to the controls.

We similarly studied a model with SNP-drug interactions selected via the stepwise method. We analyzed the predictive power of the model in terms of in-sample and out-of-sample predictions (Fig. 22). The fitting curve is above its respective controls and consequently the AUC value is higher (68.2% vs. 65.5%). The leave-one-(patient)-out cross-validation curve is overlapping with its control line (AUC~62%). Figure 22 shows that a model containing selected SNP-drug interactions has predictive power but it is nearly equivalent to randomly picking the same number of interactions from the dataset (as it has been shown in previous analyses). Altogether the SNPs, the drugs and their interactions explained 4.86% of the lipid levels variations after cross-validation (17.17% after fitting). Selecting the SNP- drug terms from the whole dataset and fixing them before running the cross-validation, the predictive performances are better (figure not shown), revealing that the set of interactions selected within the cross validation process vary from one patient to another, which was also the case for the linear terms. Adding interactions in a stepwise manner did not significantly improve the predictive performance.

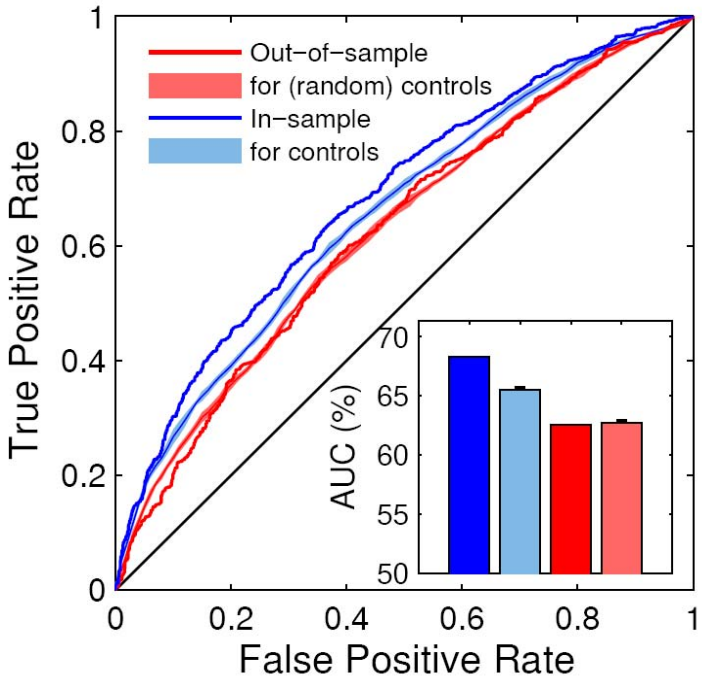


Figure 22: ROC curves and AUC values for the model containing selected SNPs, drugs and SNP-drug interactions.

On the left side are represented four ROC curves and on the right side box, the corresponding AUC values. The blue curve and bar plot refer to the in-sample predictions and AUC. The light blue ones correspond to the controls. The red curve and bar plot refer to the out-of-sample predictions and AUC. The light red ones correspond to the controls.

The same analysis has been performed using SNP-SNP interactions (Fig. 23) .The fitting curve is above its respective controls with higher AUC values (67.95% vs. 65.58%). The leave-one-(patient)-out cross-validation curve is under its control line (AUC=59.11% vs. AUC=61.23%) meaning that selecting the SNP-SNP interactions performs worse than picking them at random, reflecting the absence of trend or pattern in the SNP-SNP interaction space and consequently reflecting their lack of predictive power for the TG response.

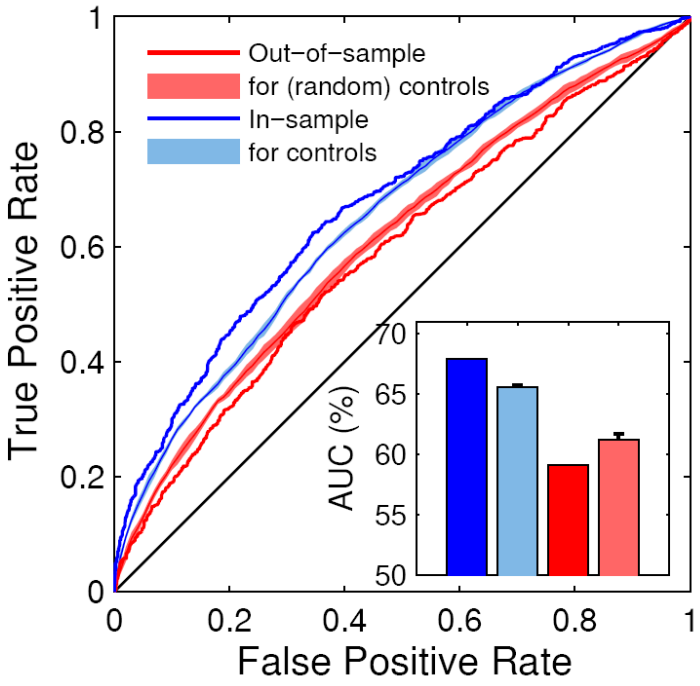


Figure 23: ROC curves and AUC values for the model containing selected SNPs, drugs and SNP-SNP interactions.

On the left side are represented four ROC curves and on the right side box, the corresponding AUC values. The blue curve and bar plot refer to the in-sample predictions and AUC. The light blue ones correspond to the controls. The red curve and bar plot refer to the out-of-sample predictions and AUC. The light red ones correspond to the controls.

Concerning the model with drug-drug interactions, the same analysis has been performed (Fig. 24). The fitting curve is above its respective controls with higher AUC values (66.14 % vs. 65.76%). The leave-one-(patient)-out cross-validation curve is under its control line (AUC=62.22% vs. AUC=63.02%) meaning that selecting the drug-drug interactions performs worse than picking them at random, likely due to the same reasons as the others interactions. The different interactions when selected before the cross validation demonstrate predictive power in out-of-sample context (figure not shown). The explained variance after cross-validation reaches the value of 7.15%

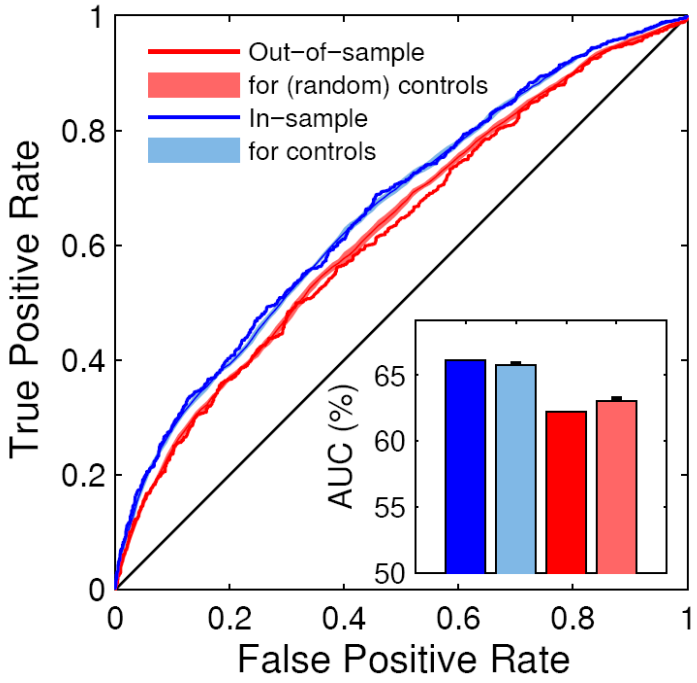


Figure 24: ROC curves and AUC values for the model containing selected SNPs, drugs and drug-drug interactions.

On the left side are represented four ROC curves and on the right side box, the corresponding AUC values. The blue curve and bar plot refer to the in-sample predictions and AUC. The light blue ones correspond to the controls. The red curve and bar plot refer to the out-of-sample predictions and AUC. The light red ones correspond to the controls.

Based on the performance of the different models evaluated, it seems that the signal is mainly driven by the drugs. The different drugs used in HAART are known to affect the lipid metabolism and transport. Therefore, we designed a model which takes into account only the SNP information (SNPs and non identical SNP-SNP interactions) (Fig. 25).

The fitting curve is always above its respective controls and consequently the AUC value is higher (~63 % vs. ~60%). The leave-one-(patient)-out cross-validation curve is overlapping with its control line (AUC~55%). Figure 25 shows that the SNP-SNP interactions combined with the SNPs do not help in reaching very high predictive performance.

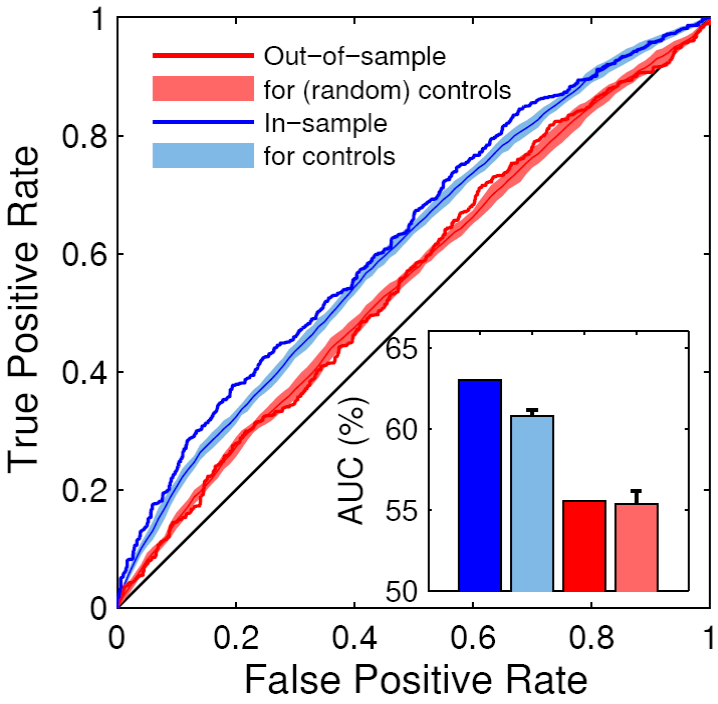


Figure 25: ROC curves and AUC values for the model containing selected SNPs and SNP-SNP interactions.

On the left side are represented four ROC curves and on the right side box, the corresponding AUC values. The blue curve and bar plot refer to the in-sample predictions and AUC. The light blue ones correspond to the controls. The red curve and bar plot refer to the out-of-sample predictions and AUC. The light red ones correspond to the controls.

	Fitting			Controls		
	AUC	r^2	R^2	AUC	r^2	R^2
SNPs	0.5858	0.0477	0.0476	0.5565+/-9E-3	0.0232+/-8.3E-3	0.0230+/-7.9E-3
Drugs	0.6103	0.0747	0.0733	0.6123+/-9.97E-4	0.0751+/-1.1E-3	0.074+/-1.2E-3
SNPs + Drugs	0.6601	0.1348	0.1346	0.6427+/-6.8E-3	0.1070+/-9.6E-3	0.1066+/-9.2E-7
SNPs + Drugs + SNP*Drug Interactions	0.682	0.1718	0.1717	0.6551+/-1.5E-3	0.1328+/-1.3E-3	0.1325+/-1.2E-3
SNPs + Drugs + SNP*SNP Interactions	0.6795	0.1806	0.1806	0.6558+/-1.5E-3	0.1365+/-2E-3	0.1364+/-1.9E-3
SNPs + Drugs + Drug*Drug Interactions	0.6614	0.1506	0.1503	0.6576+/-1.2E-3	0.1375+/-2.8E-3	0.1372+/-2.6E-3

	Leave-one-patient-out cross-validation			Controls		
	AUC	r^2	R^2	AUC	r^2	R^2
SNPs	0.548	0.0178	0.0103	0.5250+/-9.3E-3	0.0066+/-3.2E-3	-4.22E-4+/-7.5E-3
Drugs	0.5625	0.0385	0.0316	0.5999+/-3.6E-3	0.0593+/-4.5E-3	0.0581+/-4.5E-3
SNPs + Drugs	0.6237	0.083	0.0763	0.6185+/-5.1E-3	0.0737+-/7.5E-3	0.0702+/-8.6E-3
SNPs + Drugs + SNP*Drug Interactions	0.6254	0.0726	0.0486	0.627+/-1.6E-3	0.0856+/-2E-3	0.0793+/-3E-3
SNPs + Drugs + SNP*SNP Interactions	0.5911	0.045	-0.0245	0.6123+/-4.2E-3	0.066+/-4.9E-3	0.0424+/-6.6E-3
SNPs + Drugs + Drug*Drug Interactions	0.6222	0.0849	0.0715	0.6302+/-2E-3	0.0908+/-5.5E-3	0.082+/-3E-3

Table 7: Summary of AUC, r^2 and R^2 values across the models for both in-sample (fitting) and out-of-sample (leave-one-out cross-validation).

The upper table shows the AUC, r^2 and R^2 values for the different stepwise selected models after fitting and the same values for their corresponding control models. The lower table shows the AUC, r^2 and R^2 values the different stepwise selected models after a leave-one-out cross validation and the same values for the corresponding control models. AUC, r^2 and R^2 values for control models are indicated as mean+/-standard deviation.

As the triglyceride response is a complex trait, we tried to model and predict its values in HIV treated individuals with the help of their genetic variants and the drugs taken. It seems that already the SNP and the drug terms carry most of the predictive power. With the different models described above and using a *stratified* cross-validation approach, we could detect a certain predictive power for model containing only SNPs and model with SNPs and drugs (compared to models with randomly selected terms). For the other models studied, mainly models containing drug terms and interactions terms, we could not detect a significant difference with respect to control models probably due to missing factors, the small sample size, potential noise of the dataset or simply lack of large interaction effects.

2.7 General conclusions and discussion

In this project, we were interested in explaining and predicting dyslipidemia, a major lipid disorder, in treated HIV-infected individuals. We therefore investigated the different features from the dataset in order to build statistical models taking into account the genetic factors and drug treatment. In particular, we searched for possible interactions between SNPs and drugs that could explain some of the variation in lipid levels. Finally, we assessed the predictive powers of the different models.

We compared the performance of different approaches for integrating SNP-drug interactions into a linear model in a systematic manner. We established a new procedure that first rates each interaction individually and then builds a more comprehensive model using only the most significant interactions. The more established step-forward (or stepwise) approach is computationally more expensive than the two-stage approach, while its gain in R^2 and AUC for our data was not dramatic. The fact that the stepwise procedure gave better predictive performances may be because this approach is designed to select the best interaction at each step (local optimum), thus avoiding redundant interactions to be part of the model, which could have happened in the two-stage approach.

We showed that the procedures for establishing linear models with a biased selection of SNP-drug interactions have better *explanatory* power than the controls with random selections. Yet, we could not robustly establish that these procedures also have significantly better predictive power in a *stratified* cross-validation (where data from one patient are kept in the same fold). The only model exhibiting better predictive power when comparing to control models in terms of cross-validation was the one containing only SNP terms. The predictive power in terms of AUC of a model containing only SNPs and drugs is nearly equivalent to that of a model having also SNP-drug interactions, irrespective of whether they were selected by stepwise model selection or at random. In addition, the explained variance actually decreases when adding interaction terms into the model. This result is probably due to data overfitting, the small sample size and the potential noise of the dataset. Yet, it is also plausible that the data we investigated simply have no effects or that there is too small signal to allow detection of SNP-drug interactions. One should not forget that there was poor gene coverage and that the statistical methods used assume allelic homogeneity. If multiple rare variants in

each gene play a role, these methods may not be able to capture the underlying signal. In this project, we only studied interactions with a limited set of candidate SNPs, which may not tag variants that mediate the relevant SNP-drug interactions. Some genetic variants in other genes or non-recorded confounding factors (such as alcohol intake not recorded in the SHCS) could also play a role in the variation of lipid levels. In order to avoid spurious results, we adjusted the TG phenotype for various factors present in the data, which could influence the lipid levels, such as patient id, sex, age, fasting state, LLA, DM2, BMI, smoking status, ethnicity and waist. We also analyzed only the Caucasian individuals (~85%) and did not find significant differences.

Concerning the explained variance, we realized that using R^2 (as defined in eq. 5) is problematic for evaluating out-of-sample prediction, because it decreased when adding interaction terms and even became sometimes negative. Moreover, when fitting models it also is not the ideal measure to indicate the optimal number of interaction terms, because even for random interactions R^2 increases as a function of the number of terms. One could use the adjusted R^2, which takes into account the number of features into the model. Or the Mean Squared Error (MSE), measure similar to R^2 since it is used to refer to residual sum of squares, divided by the number of degrees of freedom. However, the main use of the R^2 was to compare R^2 from given models with respect to R^2 from control models (both having the same number of terms) with the aim of showing that a smart selection of interactions would give better predictive performances than a random set of interactions.

In most our analyses, we used cross-validation approaches to reflect the situation where the predicted lipid measurement was not used in the training process. We mainly used leave-one-patient-out cross-validation. Indeed, using cross-validation schemes that distributed the measurements of one patient across different folds (*random* cross-validation) indicated significant predictive power. Yet we believe that this result is less interesting, because in a realistic scenario one will have to predict lipid responses for new patients for which no measurements will be available. The marked difference between our results for *stratified* and *random* cross-validation are likely due to the fact that the lipid measurements stemming from the same patient are often correlated, while our models largely ignore this correlation. We could have averaged out measurements from the same individuals when the same drug was given, but this would have meant less phenotypic measurements available for the modeling

approaches. Another option to improve predictive power would have been to weight consistent measurements from the same patients higher.

While some models ignore the inter-dependency of measurements taken on the same patient, future work should investigate whether proper modeling of these correlations would improve the predictive power. Indeed, there is a class of models that include these types of effects, the "random effects" in mixed models [22]. We nevertheless showed that a model containing only SNP terms was able to better predict the triglyceride levels than a model built from randomly selected SNPs. Since SNP patterns reflect the patient identity, this positive result may be simply due to patient correlation rather than a direct effect of the genetic variants on lipid metabolism, while this "artificial" predictive signal is diluted when also including drugs and interactions in the model.

Further work is needed to address the predictive power of SNP-drug interactions. We cannot rule out that more sophisticated model selection procedures (like LASSO [65] or ridge regression [23]) would yield better predictive power than the procedures we investigated. It is also plausible that going beyond linear modeling is necessary to demonstrate increased predictive power when including interactions. This includes methods for classification like Neural Networks [16] or Support Vector Machines [66].

We believe that our report of the challenges we encountered in detecting SNP-drug interactions in the context of the lipid responses in the Swiss HIV Cohort Study will be useful for future work in this field. We hope that our work will provide important lessons also for other computational projects directed towards the goal of "personalized" medicine by integrating gene-environment interactions in a data-driven analysis. This is particularly important since the amount of data generated by new technologies such as sequencing will raise new challenges and there will be an increase need of statistical approaches in order to reduce the complexity of the data and capture relevant signatures like gene-environment interactions.

3 Genome-Wide Association Studies projects

During the second half of my PhD, I was also involved in different projects related to genome-wide association studies (GWAS). In contrast to my HIV project, all these projects employed genotypic data (usually Single Nucleotide Polymorphisms, generated using SNP arrays) across the entire genome.

3.1 Genome-Wide Association Studies: Benefits and current limitations

The goal of these studies was to search for correlations between genetic markers and any measurable trait in a population of individuals. The motivation is that such associations could provide new candidates for causal variants in genes (or their regulatory elements) that play a role in the phenotype of interest. In the clinical context, this may eventually lead to a better understanding of the genetic components of diseases and their risk factors.

Most of these GWAS usually currently include the following steps:

- genotype calling from the raw chip-data and basic quality control
- principle component analysis (PCA) to detect possible population stratification
- genotype imputation (using linkage disequilibrium information from HapMap)
- statistical testing for association between a single genetic marker and continuous or categorical phenotypes
- correction for multiple hypothesis testing
- data presentation (e.g. using quantile-quantile (*p*-value inflation) and Manhattan plots (*p*-value values as a function of chromosomal position))
- cross-replication and meta-analysis for integration of association data from multiple studies. Replication of findings in different studies can strengthen the evidence for a real association. By combining the results of different studies, statistical meta-analysis can also provide additional power for the discovery of new associations.

The first wave of large-scale, high-density genome-wide association (GWA) studies has improved our understanding of the genetic basis of many complex traits [67]. For several diseases, including hypertension [68, 69], type 1 [70] and type 2 diabetes [71, 72], inflammatory bowel disease, prostate cancer [73] and breast cancer [74], there has been a rapid increase in the number of loci implicated in these diseases. For others, such as asthma

[75], coronary heart disease [76], fewer novel loci have been found, although opportunities for mechanistic insights are equally promising. Several common variants influencing important continuous traits, such as lipid traits [71, 77-79], height [80], body mass index (BMI) [81] , and fat mass [82] have also been found. A recent catalog of GWAS hits lists 779 published loci for 148 traits (see Fig. 1, http://www.genome.gov/26525384).
Recent successes in the identification of susceptibility variants that underlie many important biomedical phenotypes has increased confidence that this information can be translated into clinically beneficial improvements in disease management [83].

From the many GWAS that were performed in recent years it has become apparent that even well-powered (meta-) studies with many thousands (and even ten-thousands) of samples can at best identify a few (dozen) candidate loci with highly significant associations. Most common variants that have been found to be associated with disease through GWAS typically have very small effects on the variability of the trait and explain a rather small portion of the heritability. These initial findings suggest that many GWAS may have not been sufficiently powered to discover associations with such small effects and therefore stimulated the creation of consortia to merge results from several GWAS in order to reach sufficient statistical power to identify smaller and smaller genetic effects. Increasing sample size indeed provides the required power, but the clinical significance of these findings remains an open question [84].

While many of these associations have been replicated in independent studies, each locus explains but a tiny (<1%) fraction of the genetic variance of the phenotype (as predicted from twin-studies). Remarkably, models that pool all significant loci into a single predictive scheme still miss out by at least one order of magnitude in explained variance. Thus, while GWAS already today provide new candidates for disease-associated genes and potential drug targets, very few of the currently identified (sets of) genotypic markers are of any practical use for accessing risk for predisposition to any of the complex diseases that have been studied. Many of the greatest challenges to be faced in the years ahead lie not so much in the identification of the association signals themselves, but in defining the molecular mechanisms through which they influence disease risk and/or phenotypic expression [83].

Various solutions to this apparent enigma have been proposed: First, it is important to realize that the expected heritabilities usually have been estimated from twin-studies, often several decades ago. It has been argued that these estimates entail problems of its own (independently

raised twins shared a common prenatal environment and may have undergone intrauterine competition, etc.).

Second, the genotypic information is still incomplete. Most analyses use microarrays probing only around half a million of SNPs, which is almost one order of magnitude less than the current estimates of about 4 million common variants from the Hapmap CEU panel. While many of these SNPs can be imputed accurately using information on linkage disequilibrium, there still remains a significant fraction of SNPs which are poorly tagged by the measured SNPs. Furthermore, rare variants with a Minor Allele Frequency (MAF) of less than 1% are poorly interrogated by these SNP-chips, but may nevertheless be the causal agents for many phenotypes. Finally, other genetic variants like Copy Number Variations (CNVs) (or even epigenetics) may also play an important role.

Third, it is important to realize that current analyses usually only employ additive models considering one SNP at a time with few, if any, co-variables, like sex, age and principle components reflecting population substructures. This obviously only covers a small set of all possible interactions between genetic variants and the environment. Even more challenging is taking into account purely genetic interactions, since already the number of all possible pairwise interactions scales like the number of genetic markers squared.

3.2 Overview of my projects

I was involved in the following five projects:

1. A meta-analysis of serum calcium across different cohorts (meta-analysis), one being the Cohorte Lausanne (CoLaus).
2. A GWAS aimed at studying the interaction between SNPs and body mass index on two-hour glucose levels (after an oral glucose test tolerance) within the MAGIC consortium.
3. A project that investigated the impact of three atrial natriuretic peptide (ANP) gene variants on HDL-cholesterol and other metabolic syndrome components in overweight/obese people. This work has been done in collaboration with the INSERM and Paris 7 University.
4. Within the GIANT consortium, I have worked with a couple of members of my group and run different GWAS on the following phenotypes: hip circumference, hip circumference controlled for BMI and overweight/obese vs. normal weight case-control studies.
5. Finally, I am involved in the Hypergenes consortium. Its aims at building a model to dissect complex genetic traits, using Essential Hypertension (EH) as a disease model with the main outcome, to find human genes responsible for EH, using a whole genome association approach. I have been doing data processing, genotype calling, quality control on the genotypic data and collection of the phenotypic measurements.

For projects 1-4, I performed GWAS on the Cohorte Lausanne sample (CoLaus). CoLaus [85] is a population-based study, which consists of a sample of individuals of Caucasian origin from Lausanne. The study includes 5435 individuals between 35 and 75 years old, of which 52.5 % are women and 47.5 % are men. Participants were extensively phenotyped (in terms of clinical and biological measurements as well as questionnaire responses related to life style and history) and were genotyped using Affymetrix Human Mapping 500K Array. The CoLaus study was approved by the Institutional Ethic's Committee of the University of Lausanne. The last project (5) used data from the Hypergenes consortium.

In the following chapters, I will describe the projects, my contribution followed by a detailed discussion.

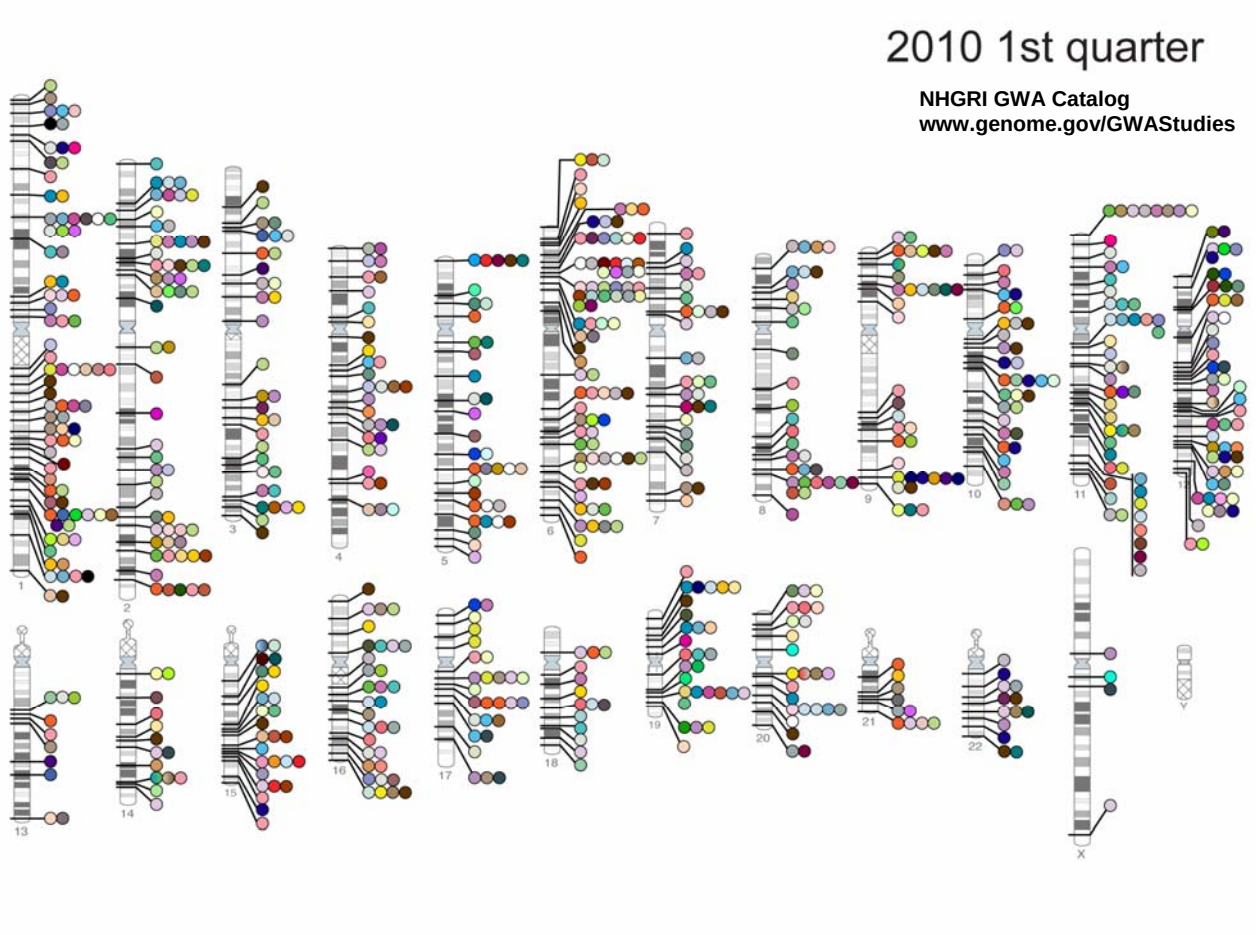


Figure 1: Published Genome-Wide Associations through March 2010, 779 published GWA at p≤5E-8 for 148 traits.

3.3 Genome-wide association for serum calcium

I have been involved in a project whose aim was to uncover genetic variations influencing human serum calcium levels at genome-wide scale. This project has been initiated in Lausanne and statistical analyses were led by Dr. Karen Kapur, a post-doc in the Computational Biology Group.

3.3.1 Background

Calcium levels in blood serum play an important role in many biological processes. As a universal cellular signaling molecule [86], calcium is involved in membrane potential, heart rate regulation and generation of nerve impulses. It also influences bone metabolism and ion transport [87]. Serum calcium levels have been associated with numerous disorders of bone and mineral metabolism as well as with cardiovascular mortality. The regulation of serum calcium is under strong genetic control, with twin studies showing that the variance in total calcium due to genetic effects is between 50% and 78%.

3.3.2 Scope of this project

This study presented the first meta-analysis of genome-wide associations from four cohorts totaling 12,865 participants of European and Indian Asian descent. It identified common polymorphisms at the calcium-sensing receptor (*CASR*) gene locus that were associated with serum calcium concentrations. We showed that *CASR* variants give rise to the strongest signals associated with serum calcium levels in both European and Indian Asian populations, while no other locus reaches genome-wide significance, indicating that *CASR* is a key player in genetic regulation of serum calcium in the adult general population.

Our analysis entitled *Genome-wide meta-analysis for serum calcium identifies significantly associated SNPs near the calcium-sensing receptor (CASR) gene*, was published in PLoS Genetics in July 2010. It can be found in Appendix 2 and also online: http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1001035.

3.3.3 My contribution

Here I describe the data analysis that I performed on serum calcium and calcium-related outcomes on the CoLaus study. My main contribution was to replicate the major results found by Karen Kapur. In order to provide an independent confirmation of her results obtained

using our analysis tool QUICKTEST [88] (that was developed in our group), I used Matlab® scripts for my analysis.

I first extracted from the CoLaus data [85] a subset of 5404 individuals for whom serum calcium measurements and questionnaire responses to queries about personal histories of osteoporosis and osteoarthritis were available. Their genotypes were either measured or imputed (prior imputation of allele dosage of SNPs was performed using IMPUTE [89] and SNPs typed in the HapMap CEU population) giving a total of ~2.5 million SNPs.

Then, I computed active serum calcium which I estimated from total serum calcium by the following formula: Ca_corrected (mmol/L) = Ca_total (mmol/L) + (40 – albumine [g/L])/40. Individuals with outliers for corrected calcium values, i.e. <1.9 or >3.0 mmol/L were excluded. I log transformed the phenotype, adjusted for age and pseudosex (pseudosex is a variable with three categories: men, premenopausal women and postmenopausal women. I defined the new phenotype used for the following association analysis using the residuals from the linear regression.

I carried out linear-regression analyses for association using an additive genetic model on the new phenotype. I corrected *p*-values for inflation using genomic control methods [90] for genotyped and imputed SNPs. I generated a text file containing my analysis results for all SNPs. It included the information in Table 1.

Column	Description
name	SNP rs number
chromosome	Chromosome number
position	physical position for the reference sequence on the Genomic build 35
strand	Strand (+/-) indicating either the positive/forward strand or the negative/reverse strand
referenceallele	Reference SNP allele (A/C/G/T)
modeledallele	Modeled (or coded) SNP allele (A/C/G/T)
build	Genomic build used
modelledallelefrq	modeled (or coded) allele frequency
n	Number of individuals with both genotype and phenotype measurements within the analyzed group
beta	effect of the modeled allele
sebeta	standard error of the effect
p	uncorrected *p*-value for the additive model
pexhwe	exact Hardy-Weinberg test *p*-value
call	genotyping call rate for the SNP, within the analyzed subgroup
imputed	1/0 coding ; 1 = imputed SNP ; 0 = directly typed SNP
usedforimputation	1/0 coding ; 1 = used for imputation ; 0 = not used for imputation
rsqhat or .info statistic	used for imputation quality assessment

Table 1: Genome-wide association results format.

Figures 1 and 2 show the Manhattan plot and quantile-quantile plot of the association results for CoLaus.

I also investigated the association of rs1801725, the top hit in the European cohorts, with several outcomes postulated to be correlated with serum calcium. I performed the analysis of osteoporosis and osteoarthritis status (binary responses) in CoLaus, using logistic regression including age and pseudosex as covariates. Table 3 reports the $-\log_{10}$ transformed *p*-value for association, the effect size and its standard error, the odds ratio and its confidence interval.

3.3.4 Results

Confirming Karen's findings, my analysis showed 23 SNPs exceeding the genome-wide significance threshold of 5E-8, all located on chromosome 3 (Fig. 1). My quantile-quantile plot of the association results is shown in Figure 2. Even though there is no excessive *p*-value inflation, Figure 2 shows deviation from the expected line and most of the signal corresponds to SNPs on chromosome 3. The two top hits for association with serum calcium (in terms of *p*-values) are rs17251221 and rs1801725. Their exact location, their coded and non-coded alleles, the beta estimates, its standard errors and the *p*-value for association are reported in Table 2. According to the NCBI SNP database, the two SNPs are located in the calcium-sensing receptor locus (CASR).

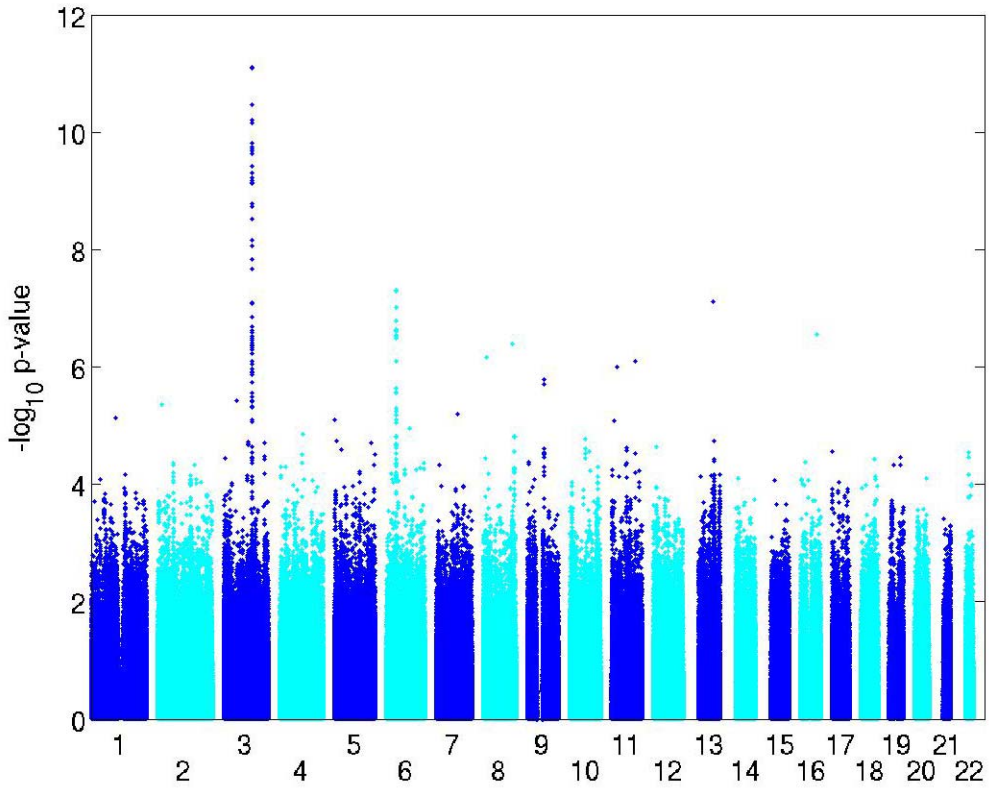


Figure 1: Manhattan plot for the Genome-wide association results for CoLaus.

The figure shows the significance of association of all SNPs in the GWAS for CoLaus after applying genomic control correction. SNPs are plotted on the x-axis according to their position on each chromosome against association with serum calcium concentrations on the y-axis (shown as $-\log_{10}$ *p*-values).

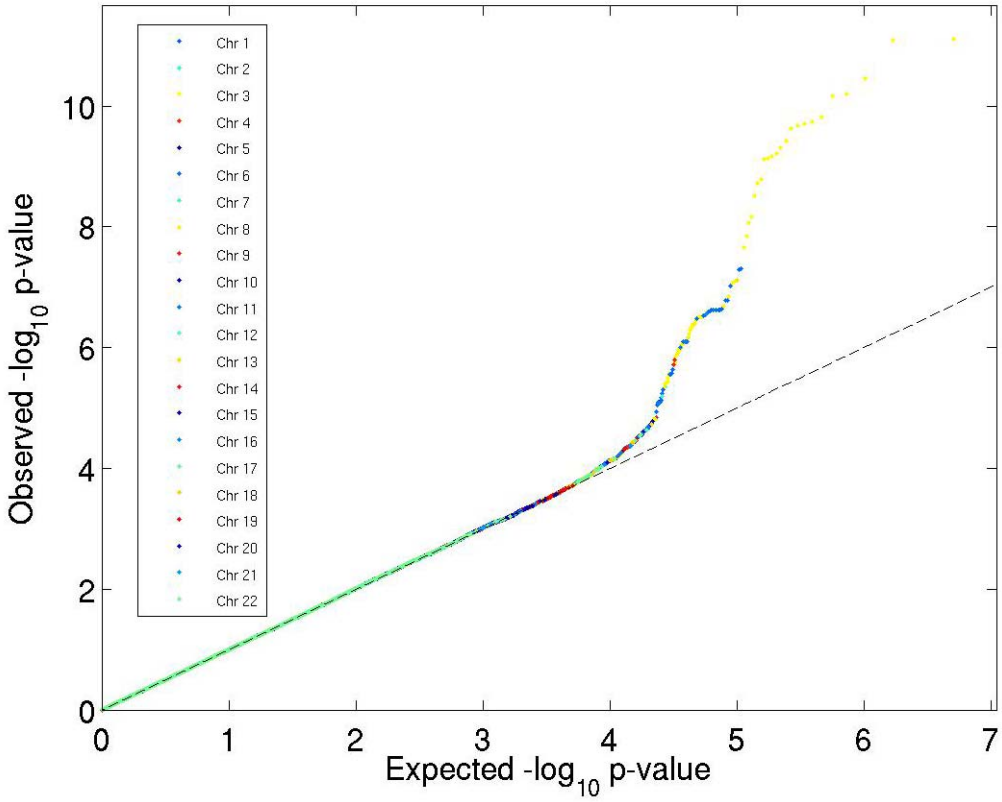


Figure 2: Quantile-Quantile plots of genome-wide association results for CoLaus.

Shown are observed -$\log_{10}$ *p*-values on the y-axis plotted against expected -$\log_{10}$ *p*-values on the x-axis after applying genomic control correction (λ = 1.0145). Results are color-coded by chromosome. The top results largely derive from the CASR locus on chromosome 3.

db SNP	Chr	Position (Build 35)	Coded Allele	Non-Coded Allele	Beta	Se	GC *P*-value
rs17251221	3	123475937	G	A	7.11E-03	1E-03	7.89E-12
rs1801725	3	123486447	T	G	7.11E-03	1E-03	8.01E-12

Table 2: Two top hits genome-wide significant.

Are reported the two top hits for association. Chr is the Chromosome number. Position is given for NCBI Build 35. The coded allele is the allele to which the beta (effect) estimate refers. Se stands for standard error of the beta coefficient. GC *p*-value is the *p*-value after genomic control correction.

I also analyzed the association of the rs1801725 and rs17251221 with two calcium-related outcomes: osteoarthritis, osteoporosis. Logistic regression including age and pseudosex as covariates showed a significant association between rs1801725 and osteoarthritis (*p*-value = 0.0281, not significant after Bonferroni correction) whereas it did not give a significant association between rs1801725 and osteoporosis (*p*-value = 0.213) (Table 3). I observed similar results and conclusions between rs17251221 and the two outcomes.

Ca-related outcomes	beta	se beta	OR	OR CI	*p*-value
Osteoporosis	0.1732	0.1391	1.18	[0.9054; 1.5618]	0.213
Osteoarthritis	-0.1466	0.0668	0.8636	[0.7577; 0.9843]	0.0281

Table 3: Logistic regression of two clinical phenotypes on rs1801725.

Are reported the effect size (beta), its standard error (se beta), the odds ratio (OR), the odds ratio confidence interval (OR CI) and the *p*-value of the rs1801725 T allele from logistic regression of the two phenotypes.

3.3.5 Conclusions

Independently analyzing the CoLaus data, I also found that the rs1801725 T allele and rs17251221 G allele were associated with higher serum calcium in the CoLaus population (Table 2); this SNP is located on the CASR locus on chromosome 3. In our meta-analysis (see article), we observed that the rs1801725 SNP had the strongest association in individuals of European descent, while for individuals of Indian Asian descent the top hit was rs17251221. The two SNPs are in strong linkage disequilibrium.

Concerning the calcium-related outcomes within the CoLaus population, the rs1801725 T allele and the rs17251221 G allele showed some association with osteoarthritis status, although not statistically significant after a Bonferroni correction. There was no association with osteoporosis. However, the evidence of association with osteoarthritis was not supported by the meta-analysis (see article).

Our analysis suggests that CASR is a key player in the regulation of serum calcium in the adult general population.

3.4 Two-hour glucose Genome-wide association studies

3.4.1 Background

MAGIC (the Meta-Analyses of Glucose and Insulin-related traits Consortium) represents a collaborative effort to combine data from multiple GWAS to identify additional loci that impact glycemic and metabolic traits.

Two-hour post-load glucose measured during the oral glucose tolerance test (OGTT) defines glucose tolerance, reflects insulin resistance and correlates with cardiovascular risk. Yet common genetic determinants of two-hour glucose are largely unknown. Identifying such variants might lead to better understanding of type 2 diabetes (TD2) pathophysiology and cardiovascular complications. T2D is defined as a state of chronic hyperglycemia defined as elevated glucose levels measured either when fasting or two-hour after glucose challenge (2-h glucose) during an OGTT [91].

Genome-wide association studies have revealed loci associated with glucose and insulin-related traits [92]. A recent meta-analysis of genome-wide association studies of fasting glycemic traits in non diabetic individuals conducted by MAGIC reported the discovery of nine new loci associated with fasting glucose (FG) (in or near *ADCY5*, *MADD*, *ADRA2A*, *CRY2*, *FADS1*, *PROX1, SLC2A2*, *GLIS3*, and *C2CD4B*) and one locus associated with fasting insulin levels (*IGF1*). The same study showed effects on FG for seven previously published glucose and/or type 2 diabetes loci *G6PC2*, *MTNR1B*, *GCK*, *DGKB*, *GCKR*, *SLC30A8*, and *TCF7L2*. They also demonstrated association of *ADCY5*, *PROX1*, *GCK*, *DGKB*, *GCKR* with type T2D [93]. This is a powerful demonstration of how analyses of continuous metabolic traits in healthy individuals can lead to the discovery of previously unsuspected T2D susceptibility genes. Detailed physiological characterization of each locus may help elucidate their role in regulation of glucose levels, insulin secretion and/or action, and identify potential pathways involved in T2D pathogenesis [92]. Another recent MAGIC meta-analysis, identified two additional novel loci (*GIPR* and *VPS13C*) associated with two-hour glucose after an OGTT [91]. *VPS13C* is the human homologue of a yeast vacuolar sorting protein and *GIPR* is the glucose-dependent insulinotropic peptide receptor that mediates the incretin effect of GIP.

3.4.2 Scope of this project

The project aims at dissecting the interaction between SNPs and body mass index (BMI) on two-hour glucose levels. Increased BMI is a risk factor for impaired glucose tolerance, and may interact with an individual's genetic susceptibility.

Therefore, the study tries to answer several questions such as: Can SNP*BMI interactions identify new variants associated with two-hour glucose not seen in the main effects? Is it possible to identify SNP*BMI interactions in known two-hour glucose loci? In order to address these questions, MAGIC is meta-analyzing the GWAS results (tests of interaction to identify new loci with heterogeneity of effect by BMI) from eight studies, totaling more than 15,000 non diabetic individuals. The consortium is following up SNPs (to test existing two-hour loci for interaction with BMI) previously associated with two-hour glucose levels (in or near *GIPR*, *VPS13C, ADCY5, GCKR, TCF7L2* genes, and in addition two SNPs in or near *HS3ST2* and *PPP1R3B*) in 15 replication samples (~ 40, 000 individuals).

3.4.3 My contribution

In this context, I have been involved in providing GWAS results on the two-hour glucose phenotype within the CoLaus study, which was subsequently meta-analyzed with the results of other cohorts by the Consortium.

Consequently, I extracted a subset of 425 individuals from the CoLaus data [85] for whom two-hour glucose measurements were available and performed two types of analyses.

Continuous BMI analyses

I fit models of the following form:

$$G_{120} = \alpha + \beta_1(\text{BMI}) + \beta_2(\text{SNP}) + \beta_c(\text{covariates}) + \varepsilon \quad (1.1)$$

$$G_{120} = \alpha + \beta_1(\text{BMI}) + \beta_2(\text{SNP}) + \beta_3(\text{BMI} * \text{SNP}) + \beta_c(\text{covariates}) + \varepsilon \quad (1.2)$$

Here G_{120} is the untransformed two-hour glucose, BMI is the untransformed Body Mass Index (weight/height^2) and the SNP is coded additively (i.e. dosage of the minor allele). The additional covariates included in the models are: age (untransformed and continuous), sex and the ten first principal components. Following the analysis protocol, I previously excluded all

individuals with diagnosed T2D (and/or using medication for diabetes), having fasting glucose >=7mmol/l and BMI <18.5kg/m^2 or missing BMI information.

For model (1.2), I performed linear regression analyses using Matlab®, and I generated a text file including (for each SNP) the information in Table 1.

Column	Description
CHR	Chromosome
POS_#	Position on NCBI build 35
SNP	rs number of the SNP
STRAND	Strand of the chromosome on HapMap (+/-)
N	Sample size for the SNP
EFFECT_ALLELE	Allele for which the effect is reported
NON_EFFECT_ALLELE	Other allele at the SNP
EFFECT_ALLELE_FREQ	Effect allele frequency
BETA_2	β_2 – the coefficient for the SNP
SE_BETA_2	s.e(β_2) – the *robust* standard error of β_2
P_BETA_2	*P*-value for SNP
BETA_3	β_3 – the coefficient for interaction
SE_BETA_3	s.e(β_3) – the *robust* standard error of β_3
P_BETA_3	*P*-value for interaction
COVAR_BETA_2_BETA_3	Cov(β_2,β_3) – the covariance between β_2 and β_3
CALL_RATE	The call rate for the SNP
R2HAT	Imputation accuracy measure between true and imputed genotypes

Table 1: Genome-wide association results format.

Similarly, for model (1.1) I produced the same information excluding the four terms related to the SNP*BMI interaction (i.e. the covariance between β2 and β3 and those related to β3).

BMI stratified analyses

For these analyses referred to as main effects models, I stratified the CoLaus individuals by BMI into three groups:

Group1: 18.5 <=BMI < 25 (N=200)

Group2: 25 <=BMI < 30 (N=155)

Group3: 30 <= BMI < 50 (N= 70)

I then considered the following model:

$$G_{120} = \alpha + \beta_1(BMI) + \beta_2(SNP) + \beta_c(covariates) + \varepsilon \quad (1.3)$$

where G_{120} is untransformed two-hour glucose, BMI is untransformed and treated continuously and the SNP is coded additively. The additional covariates included in the model are: age (untransformed and continuous), sex and the ten first principal components.
Following the analysis protocol, I excluded all individuals with diagnosed T2D (and/or using medication for diabetes), having fasting glucose >=7mmol/l and BMI <18.5kg/m^2, BMI >=50kg/m^2 or missing BMI information.

Then for the model (1.3) I performed linear regression analyses of two-hour glucose levels, independently for each BMI group again using an additive genetic model and I provided the output (as described previously) as text files format (one per BMI group).

3.4.4 Results and conclusions

The first two scans I performed for the two-hour glucose phenotype included BMI as a covariate, investigating the main effect of each SNP (model 1.1) and additionally the SNP*BMI interaction effect (model 1.2). In the third GWAS, I ran stratified analyses according to the BMI of the individuals (three groups) and searched for SNP main effects. My analyses did not reveal any genome-wide significant hit. The lack of association might be due to the small sample size (N=425). My analysis results are currently being meta-analyzed with results from eight other cohorts, aiming at detecting some SNP*BMI interactions influencing the two-hour glucose levels in up to 15,000 healthy individuals.

Genetic studies [93] of glycemic traits identified T2D risk loci, as well as loci containing genetic variants that are associated with a modest elevation in glucose levels but are not associated with overt diabetes. In-depth physiological investigation should help to further understand glucose homeostasis in humans and may reveal new pathways for diabetes therapeutics. Two-hour glucose level is a heritable quantitative trait known to be associated with diabetes. Identification of genetic variation underlying this trait and contribution to T2D in non diabetic individuals have been established recently [91].

This and other ongoing meta-analyses conducted by MAGIC may identify common genetic variants and potential SNP*BMI interactions that are specifically associated with two-hour glucose levels in diabetes free individuals. If successful, those meta-analysis results will provide novel insights into the pathogenesis of hyperglycemia, type 2 diabetes, related metabolic disorders and cardiovascular diseases.

3.5 Study of the impact of atrial natriuretic peptide gene variants on HDL-cholesterol and other metabolic syndrome components in overweight/obese people, a replication study

3.5.1 Background

According to the definition of the International Diabetes Federation (IDF), metabolic syndrome is a cluster of the most dangerous heart attack risk factors: diabetes and raised fasting plasma glucose, abdominal obesity, high cholesterol and high blood pressure. ANP is a vasoactive peptide, involved in the modulation of blood pressure and vascular tone. It seems to also play a role in lipid metabolism as it stimulates lipolysis in human adipocytes. The gene (*NPPA*) which codes for ANP therefore is a candidate for cardiovascular risk and predisposing metabolic conditions.

3.5.2 Scope of this project

This study has been carried out in collaboration with Dr. Ronan Roussel from INSERM U695, Xavier Bichat School of Medicine, University Paris Diderot, Hôpital Bichat. His group is interested in the genetics of the atrial natriuretic peptide (ANP) gene (natriuretic peptide precursor A, *NPPA*). In contrast to the three GWAS projects described previously, in the present analysis, the focus was on three SNPs (rs5063, rs5064, rs5065) in the ANP gene and their potential effects on metabolic syndrome (MS). In this context, I provided data analysis for the CoLaus cohort on these three SNPs.

Our results were subsequently meta-analyzed with raw data from six studies, totaling more than 37,800 overweight/obese individuals. A paper, entitled *Impact of atrial natriuretic peptide gene variants on HDL-cholesterol and other metabolic syndrome components in overweight/obese people* describing the discovery analysis performed on the DESIR study and the meta-analysis has been submitted to Obesity and is accessible in Appendix 3.

3.5.3 My contribution

I used the gender-specific IDF metabolic syndrome definition [94] (Table. 1) in order to classify CoLaus individuals. From the 5435 individuals, 1416 were considered as metabolic syndrome carriers, whereas 4019 were considered not to be carriers.

Metabolic syndrome	Men	Women
Waist (cm)	>=94	>=80
High Triglycerides (mmol/l)	TG>1.7 or lipidic treatment	TG>1.7 or lipidic treatment
Low HDL (mmol/l)	hdl<1.03 or lipidic treatment	hdl<1.29 or lipidic treatment
Hypertension (mm Hg)	sbp>=130 or dbp>=85 or Hyp treatment	sbp>=130 or dbp>=85 or Hyp treatment
Glycemia (mmol/l)	gluc>=5.6 or dm2	gluc>=5.6 or dm2

Table 1: Metabolic syndrome definition in men and women.

Description of the biological and clinical features defining metabolic syndrome in both male and female populations (IDF). HDL: High density lipoprotein; TG: triglycerides; sbp: systolic blood pressure; dbp: diastolic blood pressure; Hyp: hypertensive; gluc: glucose; dm2: diabetes mellitus type 2.

In addition, each component of metabolic syndrome (Table 1) was also considered individually as a phenotype used for statistical analysis of associations with the ANP gene variants. I analyzed metabolic syndrome and its constituents within the entire Caucasian CoLaus sample (n=5435) and also within the overweight population (BMI >=25 kg/m²) (n=2867) (Table 2).

Characteristics	CoLaus (overweight)
N (Men)	2867 (1619)
Age (years)	55.0± 10.7
BMI (kg/m²)	29.1 ±3.9
Waist circumference (cm)	97.9 ± 10.9
Triglycerides (mmol/l)	1.66 ± 1.36
HDL(mmol/l)	1.49 ±0.39
Metabolic syndrome (%)	44.2

Table 2: Characteristics (Mean ± SD) of overweight/obese (BMI ≥ 25 kg/m²) individuals of CoLaus.

BMI: Body mass index. HDL: High density lipoprotein. Metabolic Syndrome (MetS) was defined according to the International Diabetes Federation criteria.

The genotypic information for the three ANP SNPs of interest (rs5063, rs5064 and rs5065) had to be retrieved from our imputed CoLaus genotypes [89] since they were not directly measured on the Affymetrix 500K SNP arrays. The three SNPs are in linkage disequilibrium. Their minor allele frequencies in the overweight population are shown in Table 3.

SNP	CoLaus (overweight)	
	Minor allelic frequency (%)	Genotyping technique
rs5063 (G664A)	3.2	Imputed
rs5064 (C708T)	8.1	Imputed
rs5065 (T2238C)	13.8	Imputed

Table 3: Allelic frequencies and type of genotyping technique in the CoLaus overweight population.

I performed logistic regression association tests of the three genetic variants with metabolic syndrome and its categorical components using Matlab®. For associations of the same three variants with the continuous metabolic syndrome components, I used the standard linear regression. I tested different genetic models: additive, recessive and dominant. I ran the analyses on the overweight/obese population, stratifying individuals according to BMI, and also on the whole CoLaus population. I adjusted the phenotypes for sex, age and for BMI when considering the entire population. All the regression analyses contain an intercept, the covariates (sex, age, and possibly BMI) and the SNP of interest. The *p*-value for association, the effect size and its standard error are reported in the following tables.

3.5.4 Results

Metabolic syndrome in the overweight/obese group of the CoLaus population

The proportion of individuals having a BMI>= 25kg/m² is 52.75% (n=2867).

Logistic regression (including age and sex as covariates) showed no significant association between the variants and the metabolic syndrome phenotype in the overweight population under any of the genetic models (data not shown for the dominant model) (Table 4).

	additive model	beta	se	p
Met Syndrome	rs5063	0.0103	0.1417	0.9418
	rs5064	0.0660	0.1009	0.5131
	rs5065	6.0327e-04	0.0789	0.9939

	recessive model	beta	se	p
Met Syndrome	rs5063	0.9504	1.1733	0.4179
	rs5064	-0.0156	0.4771	0.9739
	rs5065	-0.2105	0.2852	0.4605

Table 4: Association of the ANP variants with the metabolic syndrome in the overweight/obese group of the CoLaus population, under an additive (upper part) and recessive model (lower part).
beta: the regression coefficient indicating the change per allele; se: the standard error of the corresponding beta; p: the two sided *p*-value for association.

Analysis of the metabolic syndrome components in the overweight/obese group of the CoLaus population

	additive model	beta	se	p
Waist	rs5063	-0.6317	0.6829	0.3549
	rs5064	-0.4792	0.4847	0.3229
	rs5065	-0.3828	0.3784	0.3118
TG	rs5063	-0.0094	0.0918	0.9186
	rs5064	-0.0284	0.0652	0.6629
	rs5065	0.0282	0.0509	0.5791
HDL	rs5063	-3.1525e-04	0.0246	0.9898
	rs5064	-0.0101	0.0174	0.5603
	rs5065	-0.0134	0.0136	0.3254

Table 5: Association of the ANP variants with the metabolic syndrome components in the overweight/obese group of the CoLaus population, under an additive model.
beta: the regression coefficient indicating the change per allele; se: the standard error of the corresponding beta; p: the two sided *p*-value for association.

Logistic regression (including age and sex as covariates) failed to demonstrate an association between the variants and the metabolic syndrome components in the overweight population under an additive genetic model (data not shown for the dominant model) (Table 5).

	recessive model	beta	se	p
Waist	rs5063	0.1209	5.7938	0.9834
	rs5064	0.4660	2.3087	0.8400
	rs5065	-0.4760	1.3656	0.7274
TG	**rs5063**	**2.1615**	**0.7779**	**0.0055**
	rs5064	-0.0566	0.3104	0.8553
	rs5065	-0.0706	0.1836	0.7006
HDL	rs5063	-0.2319	0.2083	0.2655
	rs5064	**0.1709**	**0.0830**	**0.0393**
	rs5065	0.0418	0.0491	0.3942

Table 6: Association of the ANP variants with the metabolic syndrome components in the overweight/obese group of the CoLaus population, under a recessive model.

beta: the regression coefficient indicating the change per allele;se: the standard error of the corresponding beta; p: the two sided *p*-value for association.

Logistic regression (including age and sex as covariates) showed significant associations between rs5063 and the triglycerides (*p*-value = 0.0055) and between rs5064 and HDL (*p*-value=0.0393), in the overweight population under a recessive genetic model (Table 6). Under the dominant model, there was no clear signal detected between the variants and the different components in the overweight population (data not shown).

Metabolic syndrome in the CoLaus population (N=5435)

Logistic regression (including age, sex and BMI as covariates) did not reveal a significant association between the variants and the metabolic syndrome phenotype in the entire population under any of the genetic models (data not shown for the dominant model) (Table 7).

	additive model	beta	se	p
Met Syndrome	rs5063	0.0402	0.1322	0.7610
	rs5064	0.1760	0.0966	0.0685
	rs5065	0.0787	0.0755	0.2970

	recessive model	beta	se	p
Met Syndrome	rs5063	0.6405	1.0152	0.5281
	rs5064	0.6178	0.4441	0.1642
	rs5065	0.1652	0.2738	0.5464

Table 7: Association of the ANP variants with the metabolic syndrome in the CoLaus population, under an additive (upper part) and recessive model (lower part).

beta: the regression coefficient indicating the change per allele; se: the standard error of the corresponding beta; p: the two sided *p*-value for association.

Analysis of the metabolic syndrome components in the CoLaus population (N=5435)

Logistic regression (including age, sex and BMI as covariates) failed to identify significant association between the variants and the metabolic syndrome components in CoLaus population under an additive genetic model (data not shown for the dominant model) (Table 8).

	additive model	beta	se	p
Waist	rs5063	-0.1173	0.2915	0.6873
	rs5064	-0.0391	0.2081	0.8511
	rs5065	0.1053	0.1626	0.5172
TG	rs5063	-0.0139	0.0560	0.8033
	rs5064	-0.0407	0.0400	0.3089
	rs5065	-0.0012	0.0312	0.9682
HDL	rs5063	0.0115	0.0185	0.5347
	rs5064	-0.0249	0.0132	0.0601
	rs5065	-0.0148	0.0103	0.1518

Table 8: Association of the ANP variants with the metabolic syndrome components in the CoLaus population, under an additive model.

beta: the regression coefficient indicating the change per allele; se: the standard error of the corresponding beta; p: the two sided *p*-value for association.

Logistic regression (including age, sex and BMI as covariates) demonstrated significant associations between rs5063 and the triglycerides (*p*-value = 0.0189) and between rs5063 and waist (*p*-value=0.0428), in the entire population under a recessive genetic model (Table 9). However, no association has been found between the variants and the different components in CoLaus under the dominant model with the exception for rs5064 with HDL (negative effect size) (Table 10).

	recessive model	beta	se	p
Waist	**rs5063**	**5.3091**	**2.6207**	**0.0428**
	rs5064	1.1216	0.9415	0.2335
	rs5065	0.5897	0.5828	0.3116
TG	**rs5063**	**1.1814**	**0.5033**	**0.0189**
	rs5064	-0.0334	0.1809	0.8534
	rs5065	-0.0115	0.1120	0.9184
HDL	rs5063	0.0898	0.1668	0.5904
	rs5064	0.0549	0.0599	0.3597
	rs5065	0.0125	0.0371	0.7357

Table 9: Association of the ANP variants with the metabolic syndrome components in the CoLaus population, under a recessive model.

beta: the regression coefficient indicating the change per allele; se: the standard error of the corresponding beta; p: the two sided *p*-value for association.

	dominant model	beta	se	p
HDL	rs5063	0.0108	0.0189	0.5679
	rs5064	**-0.0318**	**0.0142**	**0.0253**
	rs5065	-0.0196	0.0115	0.0886

Table 10: Association of the ANP variants with HDL in the CoLaus population, under a dominant model.

beta: the regression coefficient indicating the change per allele; se: the standard error of the corresponding beta; p: the two sided *p*-value for association.

3.5.5 Conclusions

From the CoLaus analysis (which served as a replication for the meta-analysis), no significant association was detected between the metabolic syndrome incidence/prevalence and the SNPs under any genetic model. This observation might be due to the way we classified the CoLaus individuals as metabolic syndrome carriers. The classification was based on the IDF definition, and not from a clinical diagnostic, which would be more reliable. However, few associations were detected under a recessive model, positively associated with TG and HDL in the overweight CoLaus population (respectively rs5063 and rs5064) (Table 6) and rs 5063 associated with increased waist and TG in CoLaus (Table 9). One significant negative association (decrease) was detected between the HDL levels and the variant rs5064 under the dominant model in the whole population, suggesting an adverse effect of this variant.

In the meta-analysis, our collaborators tested whether the metabolic syndrome and its lipid components were associated with these polymorphisms in overweight/obese participants of six cross-sectional cohorts around the world (40,000 individuals). Neither the metabolic syndrome nor waist circumference was associated with any of the three selected SNPs. However, the HDL levels were strongly positively associated with rs5065 (p-value=3E-6 and more marginally with rs5063 (p-value=0.05) and rs5064 (p-value=0.04). Triglyceride levels were negatively associated with the rare allele of rs5064, but not with the two other SNPs. Neither the meta-analysis, nor CoLaus alone replicated or confirmed the findings of the discovery analysis performed on 5212 (2248 overweight/obese) individuals from the DESIR study.

Moreover, several studies found the ANP plasmatic levels to be associated with blood lipid levels; namely, they were positively correlated with HDL and negatively with triglyceride levels. Here, our collaborators hypothesized a plausible causal link between the genetic variation rs5065 with the HDL increase, via the increase in ANP circulating levels (not directly measured in the study), in overweight populations (based on the assumption that the more the fat depot is present, the more the ANPs play they lipolytic role). HDL is in addition affected by a number of factors, including physical activity and alcohol, for which no adjustment has been made.

Another limitation of this study was that population stratification in the meta-analyses was not dealt with analytically. Several ethnic groups are represented among the cohorts. Population principal components could have been used to take this issue into account. Waist circumference could have also been used rather than BMI as a cut off to define the overweight population, since it is the main criteria of metabolic syndrome carriers. Other suggestions could be to analyze men and women separately on the different phenotypes and only combine them if there is no sex interaction. Or investigate a more direct relationship, one between plasma levels of ANP and the SNPs.

The present findings remain hypothetical and should trigger further clinical investigations to understand how ANP gene variants may affect the lipid metabolism and to test whether these effects may be relevant to diseases processes, especially in the metabolic, cardiovascular and renal fields.

3.6 Genome-wide association studies for anthropometric measures

3.6.1 Background

The GIANT (Genome-wide Investigation of ANThropometric measures) consortium is an international assembly of statisticians and genetic epidemiologists, which was established to pool genome-wide association (GWA) results on anthropometric parameters used as measures of obesity such as adult height, body weight, body mass index (BMI), waist circumference (WC) and hip circumference (HC).

Obesity is a risk factor for developing several diseases, including type 2 diabetes, cardiovascular diseases and certain types of cancer. Obesity, defined as a body mass index (BMI; in kg/m^2) of ≥30, results from an imbalance in energy intake and energy expenditure [95]. However, the underlying mechanisms are largely unknown and are being intensively studied. Recently, it has become clear that not only the amount of body fat but also its distribution is important in determining disease risk: an increasing WC as a measure of abdominal obesity is related to increased chronic disease risk and mortality, independent of BMI as a measure of general obesity [96].

Although environmental factors play an important role in the development of obesity, multiple twin and family studies have indicated that genetic factors also make a significant contribution to its etiology [97]. Many genetic loci have been identified as being associated with obesity; however, these loci only explain a small part of the genetic variance underlying the development of obesity [98, 99]. Recently, GWAS have expanded the number of genetic susceptibility loci for obesity by identifying several new single nucleotide polymorphisms (SNPs) consistently associated with both BMI and weight, and thus, contributing to obesity risk [81, 100]. The loci identified are located in or near the genes *FTO, MC4R, TMEM18, GNPDA2, SH2B1, KCTD15, MTCH2, NEGR1, BDNF*, and *ETV5*. These loci are likely to be involved in many biological pathways because they are expressed in numerous tissues. Notably, most of the new obesity genes (except *MTCH2, NEGR1*, and *ETV5*) are expressed in the hypothalamus, a crucial center for energy balance and regulation of food intake [95]. Another meta-analysis searched for genetic loci influencing obesity and fat distribution and identified three loci (near *TFAP2B*, *MSRA*, *LYPLAL1*) implicated in the regulation of human adiposity , defined by WC and waist–hip ratio (WHR)) [101].

3.6.2 Scope of this project

The specific aims of GIANT are (i) integrating the study-specific GWAS via a uniform protocol using imputed genotypic data to pool GWA results across studies, (ii) the replication of top hits, and (iii) performing the follow-up studies on these replicated hits. This consortium was already successful in describing new genetic loci for height [80] and new polymorphisms of the MC4R gene [81, 100] having impact on BMI and obesity in general.

The GIANT consortium aims towards a better understanding of the genetic contribution to obesity. This includes meta-analyzing HC, extremes phenotypes (BMI, height, and WHR controlled for BMI), as well as case-control studies for overweight/obese vs. normal weight. This provides an opportunity to explore whether extreme phenotypes are determined by other genetic loci than those associated with traits across the whole distribution.

A paper, meta-analyzing the GWAS results from various studies, entitled *Association analyses of 249,796 individuals reveal eighteen new loci associated with body mass index*, was published in Nature Genetics in October 2010. The manuscript is accessible in Appendix 4 and also online: http://www.nature.com/ng/journal/v42/n11/full/ng.686.html.

3.6.3 My contribution

In this context, I have been involved providing GWAS results on the following phenotypes: HC, HC controlled for BMI, and overweight/obesity case-control studies within the CoLaus study [85], which were subsequently meta-analyzed with other cohorts results by the Consortium. For my analyses, I extracted 5404 HC measurements and 5434 BMI measurements and performed the following analyses.

HC and HC controlled for BMI

To these data, I first fitted models stratified by gender of the following form:

$$HC = \alpha + \beta_1(age) + \beta_2(age^2) + \beta_i(PC_i) + \varepsilon \quad (1.1)$$

$$HC\_BMI = \alpha + \beta_1(age) + \beta_2(age^2) + \beta_3(BMI) + \beta_i\,(PC_i) + \varepsilon \quad (1.2)$$

Here HC is the untransformed HC measurement, BMI is the untransformed Body Mass Index (weight/height²). The covariates included in the models are: age (untransformed and continuous), age² and the ten first principal components (PC_i).

I then defined the corrected HC and HC_BMI as the inverse normal transformed residuals from linear regressions in eq. (1.1) and (1.2). For the following SNP association analysis, I regressed each SNP coded additively ((i.e. dosage of the minor allele) onto the corrected phenotypes in men and women separately using Matlab®, and I generated a text file including (for each SNP) the information in Table 1:

Column	Description
MARKER NAME	rs id of the marker analyzed
STRAND	Strand on which the alleles are reported (+/-)
N	The effective number of subjects analyzed
EFFECT_ALLELE	The allele associated with phenotypic traits
OTHER_ALLELE	Indicating the other (non-effect) allele
EAF	Effect allele frequency
IMPUTATION	Imputation posterior probability for imputed SNPs, or the integer value directly genotyped markers
R2HAT	Imputation accuracy measure between true and imputed genotypes
BETA	The regression coefficient indicating change per effect_allele
SE	The standard error of beta
P	The two-sided *P*-value for the association

Table 1: Genome-wide association results format.

Overweight/obese vs. normal weight

For these analyses, I stratified the CoLaus individuals by BMI into four groups (see also Table 2):

1. Overweight (BMI≥25 kg/m2 vs. normal-weight [<25 kg/m2]) – sex stratified
2. Obesity class I (BMI≥30kg/m2 vs. normal-weight [<25 kg/m2]) – sex stratified

3. Obesity class II (BMI≥35 kg/m2 vs. normal-weight [<25 kg/m2]) – sex pooled
4. Obesity class III (BMI≥40 kg/m2 vs. normal-weight [<25 kg/m2]) – sex pooled

N	CONTROLS (M/W)	CASES (M/W)	TOTAL (M/W)
Overweight	2567 (941/1626)	2867 (1619/1248)	5434 (2560/2874)
Obesity class I	2567 (941/1626)	869 (444/425)	3436 (1385/2051)
Obesity class II	2567	224	2791
Obesity class III	2567	56	2623

Table 2: CoLaus individuals stratified by BMI. (N: Number of individuals; M: Men; W: Women)

I then considered the following model:

$$\text{Overweight/obesity vs. normal weight} = \alpha + \beta_1(\text{age}) + \beta_2(\text{age}^2) + \beta i(\text{PCi}) + \beta 3(\text{SNP}) + \varepsilon \qquad (1.3)$$

Here the phenotype is a binary variable and the SNP is coded additively (i.e. dosage of the minor allele). The covariates included in the models are: age (untransformed and continuous), age² and the ten first principal components.

I then performed for the model (1.3) logistic regression estimating the risk of being a case predicted by SNP (additive model), independently for each BMI group and I provided the output (as described previously) and additional information (see below) as text files format.

OR: The odds ratio per effect_allele
CI_lower: The lower 95% confidence limit of OR
CI_upper: The upper 95% confidence limit of OR
P: The two-sided *p*-value for the association

3.6.4 Results and conclusions

My GWAS results were meta-analyzed with those from 45 other cohorts (more than 240,000 individuals). Specifically, our collaborators examined associations between BMI and ~2.8 million SNPs in 123,865 individuals, with targeted follow-up of 42 SNPs in 125,931

additional individuals. From the 32 significant loci, they confirmed 14 known obesity-susceptibility loci and identified 18 new loci associated with BMI ($p<5E-8$), one of which includes a copy number variant near GPRC5. Some loci map near key hypothalamic regulators of energy balance, or near GIPR, an incretin receptor.

The 32 confirmed associations found by GIANT, included 19 loci with $p < 5E-8$, 12 additional novel loci near *RBJ/ADCY3/POMC, QPCTL/GIPR, FANCL, SLC39A8, TMEM160, CADM2, LRP1B, MTIF3/GTF3A, ZNF608, PTBP2, RPL27A/TUB, NUDT3/HMGA1,* and one locus (*NRXN3*) previously associated with waist circumference. These 19 loci included all ten loci from previous GWA studies [81, 100] of BMI, two loci previously associated with body weight (*FAIM2* and *SEC16B*) and one locus previously associated with waist circumference (near *TFAP2B*). The remaining six loci, near *GPRC5B, MAP2K5/LBXCOR1, TNNI3K, LRRN6C, FLJ35779/HMGCR*, and *PRKD1*, have not previously been associated with BMI or other obesity-related traits.

Our study increased the number of loci robustly associated with BMI from 10 to 32. Together, the 32 confirmed BMI loci explained 1.45% of the inter-individual variation in BMI (estimated heritability of 40-70%), with the *FTO* SNP accounting for the largest proportion of the variance (0.34%). This suggests the existence of other genetic markers influencing BMI and obesity and/or interactions of these markers with other loci or environmental variables.

3.7 The Hypergenes Project

3.7.1 Background

Hypergenes is an EU-funded collaborative project. It includes research activities in the areas of population genetics, molecular epidemiology, clinical sciences, bioinformatics and health information technology, with predicted outcomes in the fields of prevention, early and clinical diagnosis and treatment, in addition to increasing the knowledge about the etiology of Essential Hypertension (EH) and Target Organ Damages (TOD).

Essential hypertension (also called primary or idiopathic hypertension) is the form of hypertension that by definition has no identifiable cause. It is the most common type of hypertension, affecting 95% of hypertensive patients [102]. EH refers to a persisting elevated blood pressure (BP) with heterogeneous genetic and environmental causes. Prevalence increases with age. In developed and developing countries alike, EH affects 25–35% of the entire adult population and up to 50% of those beyond the sixth decade of life. EH is a powerful risk factor for cardiovascular disease (CVD), the most common cause of morbidity and mortality in Europe [103]. Besides being a major risk factor for coronary heart disease and renal failure, EH is causally involved in nearly 70% of all strokes, the third commonest cause of death worldwide after ischemic heart disease and all types of cancer combined. By 2020, stroke mortality will have almost doubled. EH contributes pathogenetically to the development of both systolic and diastolic heart failure. EH has a strong genetic component (from 30% to up to 70% depending to the assumed mode of inheritance [68, 104]), which is modulated by several, even though not fully understood [105] environmental influences. It is likely that both genetic and non-genetic determinants of hypertension interact differently in different populations.

To date, 12 GWAS on blood pressure and hypertension have been published, mostly on participants of European origin. Only two of the published studies on blood pressure traits (CHARGE BP and Global BP Gen) have identified an association withstanding correction for multiple testing genome-wide significance. In total, 14 independent loci have been identified so far for blood pressure traits that reached genome-wide significance, including replication in independent cohorts [69, 106]. Global BP Gen (n = 34,433) identified association between systolic (SBP) or diastolic blood pressure (DBP) and common variants in eight regions near the *CYP17A1* (p = 7E-24), *CYP1A2* (p = 1E-23), *FGF5* (p = 1E-21), *SH2B3* (p = 3E-18),

MTHFR (p = 2E-13), *c10orf107* (p = 1E-9), *ZNF652* (p = 5E-9) and *PLCD3* (p = 1E-8) genes. The CHARGE Consortium (n = 29,136) identified 13 SNPs for SBP, 20 for DBP and 10 for hypertension at P < 4E-7. The top ten loci for SBP and DBP were incorporated into a risk score; mean BP and prevalence of hypertension increased in relation to the number of risk alleles carried. When ten CHARGE SNPs for each trait were included in a joint meta-analysis with the Global BP Gen Consortium (n = 34,433), four CHARGE loci attained genome-wide significance (p < 5E-8) for SBP (*ATP2B1, CYP17A1, PLEKHA7, SH2B3*), six for DBP (*ATP2B1, CACNB2, CSK-ULK3, SH2B3, TBX3-TBX5, ULK4*) and one for hypertension (*ATP2B1*).

The common variants associated with blood pressure phenotypes have a very small effect size. The significant variants so far explain only a very small fraction of the heritability of blood pressure traits. The effect sizes of the variants identified are small and currently explain about 1% of the phenotypic variability (after correcting for major confounders such as sex, age, and body mass index). Rare variants may explain more of the phenotypic variability, it is also possible that gene-environment interactions play an important role, but it is currently not possible to quantify them [68].

3.7.2 Scope of this project

The project is focused on the definition of a comprehensive genetic epidemiological model of complex traits like Essential Hypertension (EH) and intermediate phenotypes of hypertension dependent/associated Target Organ Damages (TOD). To identify the common genetic variants relevant for the pathogenesis of EH and TODs, Hypergenes performed a Genome-Wide Association Study (GWAS) of 4,000 subjects recruited from historical well-characterized European cohorts. Genotyping was done with the Illumina Human 1M BeadChip. Well-established multivariate techniques and innovative genomic analyses through machine learning techniques have been used for the GWAS investigations. Using a machine learning approach, they aim at developing a disease model of EH integrating the available information on EH and TOD with relevant validated pathways and genetic/environmental information to mimic the clinician's recognition pattern of EH /TOD and their causes in an individual patient. The experimental design included the distinct data generation for two datasets in parallel, each with around 1,000 cases and 1,000 controls, followed by a reciprocal

replication andjoint analysis. This design is more powerful than replication alone and allows also a formal testing of the potential heterogeneity of findings compared to a single step (one large sample) design. Association results will be used to build a customized and inexpensive genetic diagnostic chip that can be validated in existing cohorts (n=12,000 subjects). Designing a comprehensive genetic epidemiological model of complex traits will also help to translate the genetic findings into improved diagnostic accuracy and new strategies for early detection, prevention and eventually personalized treatment of a complex trait.

Hypergenes's Scientific and Technological Objectives and expected outcomes are (i) finding genes responsible for EH and TOD, using a whole genome association/entropy-based approach, (ii) developing an integrated disease model, considering the environment, using an advanced bioinformatics approach and (iii) testing the predictive ability of the model to identify individuals at risk.

3.7.3 My contribution and results

In contrast to the four other GWAS collaborations, in this project I was involved in one of the first step of a GWAS, which is the genotype calling from the raw chip-data and basic quality control (see also section 3.1). For Hypergenes, there were two genotyping centers, one in Milan, the other in Geneva. I performed the genotype calling corresponding roughly to half of the total sample size, including almost 2000 samples hybridized on Human 1M-Duo v2.0 BeadChips and scanned with the Illumina iScan® at the Plateforme génomique-NCCR Frontiers in Genetics, at the University of Geneva.

These chips interrogate more than 1,100,000 evenly distributed SNPs per sample. The Human1M-Duo BeadChip focuses on tag SNPs produced by the International HapMap Project, on SNPs in functional gene regions, on SNPs and non-polymorphic markers in known and novel copy number regions (CNV), on SNPs mapping in Absorption, Distribution, Metabolism and Excretion (ADME) related genes as well as Major Histocompatibility Complex (MHC) regions. SNPs in 1M BeadChip have 1.5kb median marker spacing.
After whole genome amplification, fragmentation of samples genomic DNA, and locus-specifically hybridized to each individual target on the beads, each SNP was characterized through a single-base enzymatic extension assay using two colors labeled nucleotides. After the extension, the labels were visualized by staining with a sandwich-based immunohistochemistry assay that increases the overall sensitivity of the assay.

The protocol ended with the chip scanning process and storage of fluorescence intensities and other related scan files. Fluorescence intensities were then extracted and normalized using a proprietary normalization method from Illumina's software, Genome Studio, and a Genome Studio Project (.bsc file) wad created. Clusters were created for each SNP against a cluster reference file (created using a diverse set of over 100 samples from different HapMap populations) and genotypes were attributed. Genome Studio calculated sample and SNP statistics. Genome Studio calculated the Call Rate/individual and allowed to evaluate sex mismatch. The call rate/individual is the percentage of SNPs correctly called over the total number of SNPs genotyped. We set a call rate threshold at 0.95 and only samples with call above this threshold were used for further analysis of association.

My genotyping activity is summarized in Table 1, where the number of samples genotyped divided by cohort and the number of samples with call rate above or below 0.95 are reported.

Cohort ID	Phenotype	Number of samples genotyped	Call rate >0.95	Call rate <0.95
HCS	Cases	359	354	5
WHS	Cases	78	77	1
SOP	Cases	494	457	37
IMM	Cases	164	161	3
HSH	Cases	56	56	-
LVM	Cases	181	174	7
Total Cases		**1332**	**1279**	**53**
SSA	Controls	423	401	24
WHS	Controls	46	46	-
IMM	Controls	106	101	5
IMA	Controls	88	86	2
Total Controls		**663**	**634**	**31**

Table 1: Overview of the subjects genotyped with call rate <0.95 and call rate >0.95.

Genome Studio allows to estimate the sex of the genotyped subjects, based on the SNPs mapping on sex chromosomes. This was a very useful tool to check the concordance between sex estimated on genetic data and sex reported in files sent by the Clinical centers. If sex inconsistencies were observed and confirmed after having re-checked with the Clinical centers, the DNA sample was removed from further statistical analysis of association.

Gender estimation and gender check in the sample genotyped in the Geneva center gave the results described in Table 2.

Cohort ID	Phenotype	Number of samples with sex mismatch
HCS	Cases	7
LVM	Cases	4
SOP	Cases	16
SSA	Controls	7
WHS	Control	1
SSA	Case	1
Total		**36**

Table 2: Overview of sex mismatches detected.

For each DNA sample, a final report was generated from the .bsc file. This report is the final output of Genome Studio genotyping. It is a .csv file and contains genotyping information as follows:

#	Parameter	Description
1	Sample ID	Unique identification code of the DNA
2	SNP Name	Unique identification code of the SNP
3	ILMN strand	Design strand designation
4	Customer strand	Customer strand designation
5	Chr	Chromosome number of the SNP
6	Position	SNP position on chromosome
7	Allele1 - Top	Illumina's allele definition that let you code the allele without ambiguity and changes due to the genome build you're referring to: 1. NOT-ambiguous SNPs. like A/C.G or T/C.G. Allele1 is always A or T. but A is Allele1-TOP and T is Allele1-BOT (Bottom). 2. In the case of ambiguous SNPs. like A/T and C/G Illumina's method goes reading -n-bases upstream and +n-bases downstream the SNP position up to an unambiguous situation. When such a condition is found. the nomenclature is referred to the NOT-ambiguous SNP calling
8	Allele2 - Top	
9	GC Score	GenCall score per single SNP is primarily designed to rank and filter out failed genotypes. The sensitive region of the GenCall Score is between the values of 0.2 and 0.7. Scores below 0.2 generally indicate failed genotypes. while scores above 0.7 usually report well-behaving genotypes
10	R	Normalized R-value (also named NormR) of a SNP for the sample. R-value is the intensity of the fluorescence signal
11	X	Normalized intensity of the A allele. referred to the A vs B format
12	Y	Normalized intensity of the B allele. referred to the A vs B format
13	X Raw	Raw intensity of the A allele. referred to the A vs B format
14	Y Raw	Raw intensity of the B allele. referred to the A vs B format
15	Log R Ratio	Base-2 log ratio of observed R for a SNP divided by the expected R
16	B Allele Freq	Theta-value for a SNP corrected for cluster positions. Cluster positions are generated from a large set of normal individuals. B allele frequency indicates the SNP allele composition and is linearly interpolated between 0 and 1. When SNPs are well clustered. B allele frequency values are grouped around 0. 0.5 or 1. meaning AA. AB or BB respectively

Table 3: List of parameters included in the final reports generated by Genome Studio, with a brief description of their meaning.

Raws 7 (Allele1-Top) and raw 8 (Allele2-Top) are the base (nucleotide) call for allele1 and allele 2 and represent the genotype at a specific SNP. The GC score (raw 9) is an additional quality parameter.

3.7.4 Conclusions

95.7% of the samples had a genotyping call rate above 0.95. The same procedure was performed for the samples genotyped in Milan. In this case, 96.8% of the samples had a call rate above 0.95. In total, all genotyping data of the 4061 individuals genotyped were uploaded successfully into Hypergenes servers.

The main goal of Hypergenes project is to detect new loci involved in the blood pressure traits. Genome-wide association study was therefore carried out using logistic regression analysis. The normotensive were healthy subjects above 55 years of age who had been followed for a long period with no hypertension detected and the hypertensive were subjects with hypertension detected before age of 50 or under hypertensive treatment. From the discovery phase, no genome-wide significant association signal has been found within Hypergenes. A second stage has been initiated, where 15,000 SNPs have been carried forward to replication in around 12,000 individuals. Genotyping should be finished in few weeks. At the same time, chromosome Y and mitochondrial haplogroup clustering from the discovery phase are investigated in order to identify male/female migration patterns in Europe.

Hypergenes data have also been involved in different ongoing GWAS projects. Hypergenes served as a replication cohort in GIANT consortium for height and BMI phenotypes. The latter has been described in section 3.6. Hypergenes is also among the discovery cohorts for sleeping and resting hearth rate meta-analysis. The replication phase is in progress.

References

1. Lander, E.S., *Array of hope.* Nat Genet, 1999. **21**(1 Suppl): p. 3-4.
2. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
3. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
4. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
5. Arsanious, A., G.A. Bjarnason, and G.M. Yousef, *From bench to bedside: current and future applications of molecular profiling in renal cell carcinoma.* Mol Cancer, 2009. **8**: p. 20.
6. Gruvberger-Saal, S.K., et al., *Microarrays in breast cancer research and clinical practice--the future lies ahead.* Endocr Relat Cancer, 2006. **13**(4): p. 1017-31.
7. Kroese, M., R.L. Zimmern, and S.E. Pinder, *HER2 status in breast cancer--an example of pharmacogenetic testing.* J R Soc Med, 2007. **100**(7): p. 326-9.
8. Paluszczak, J. and W. Baer-Dubowska, *Epigenetic diagnostics of cancer--the application of DNA methylation markers.* J Appl Genet, 2006. **47**(4): p. 365-75.
9. Lindblad-Toh, K., et al., *Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse.* Nat Genet, 2000. **24**(4): p. 381-6.
10. Sellick, G.S., et al., *Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays.* Nucleic Acids Res, 2004. **32**(20): p. e164.
11. Ball, S. and N. Borman, *Pharmacogenetics and drug metabolism.* Nature Biotech, 1998. **16 suppl.**(2s): p. 4.
12. Housman, D. and F.D. Ledley, *Why pharmacogenomics? Why now?* . Nature Biotech, 1998. **16 suppl.**(2s): p. 2.
13. Hewett, M., et al., *PharmGKB: the Pharmacogenetics Knowledge Base.* Nucleic Acids Res, 2002. **30**(1): p. 163-5.
14. Nicholson, J.K., *Global systems biology, personalized medicine and molecular epidemiology.* Mol Syst Biol, 2006. **2**: p. 52.
15. Yang, X., et al., *Phosphoinositide signalling links O-GlcNAc transferase to insulin resistance.* Nature, 2008. **451**(7181): p. 964-9.
16. Dolgobrodov, S.G., et al., *Artificial Neural Network: Predicted vs. Observed Survival in Patients with Colonic Cancer.* Dis Colon Rectum, 2006.
17. Shastry, B.S., *Pharmacogenetics and the concept of individualized medicine.* Pharmacogenomics J, 2006. **6**(1): p. 16-21.
18. Jeffrey, S.S., et al., *Expression array technology in the diagnosis and treatment of breast cancer.* Mol Interv, 2002. **2**(2): p. 101-9.
19. Oshita, F., et al., *Genome-wide cDNA microarray screening of genes related to the benefits of paclitaxel and irinotecan chemotherapy in patients with advanced non-small cell lung cancer.* J Exp Ther Oncol, 2006. **6**(1): p. 49-53.
20. Jain, K.K., *Applications of biochip and microarray systems in pharmacogenomics.* Pharmacogenomics, 2000. **1**(3): p. 289-307.
21. Crowther, D.J., *Applications of microarrays in the pharmaceutical industry.* Curr Opin Pharmacol, 2002. **2**(5): p. 551-4.
22. Venables, W.N. and B.D. Ripley, *Modern Applied Statistics with S*. 4th ed. 2003: Springer.

23. Hoerl, A.E., R.W. Kennard, and R.W. Hoerl, *Practical Use of Ridge Regression: A Challenge Met.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1985. **34**(2): p. 114-120.
24. Nelder, J.A. and R.W.M. Wedderburn, *Generalized Linear Models.* Journal of the Royal Statistical Society. Series A (General), 1972. **135**(3): p. 370-384.
25. Donoho, D. and J. Jin, *Higher criticism thresholding: Optimal feature selection when useful features are rare and weak.* Proc Natl Acad Sci U S A, 2008. **105**(39): p. 14790-5.
26. Coffin, J., et al., *Human immunodeficiency viruses.* Science, 1986. **232**(4751): p. 697.
27. Coffin, J., et al., *What to call the AIDS virus?* Nature, 1986. **321**(6065): p. 10.
28. UNAIDS, *Overview of the global AIDS epidemic*. 2006.
29. Palella, F.J., Jr., et al., *Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators.* N Engl J Med, 1998. **338**(13): p. 853-60.
30. Chene, G., et al., *Prognostic importance of initial response in HIV-1 infected patients starting potent antiretroviral therapy: analysis of prospective studies.* Lancet, 2003. **362**(9385): p. 679-86.
31. Tozzi, V., *Pharmacogenetics of antiretrovirals.* Antiviral Res. **85**(1): p. 190-200.
32. Montessori, V., et al., *Adverse effects of antiretroviral therapy for HIV infection.* Cmaj, 2004. **170**(2): p. 229-38.
33. Saitoh, A., et al., *Myelomeningocele in an infant with intrauterine exposure to efavirenz.* J Perinatol, 2005. **25**(8): p. 555-6.
34. Lazarou, J., B.H. Pomeranz, and P.N. Corey, *Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies.* Jama, 1998. **279**(15): p. 1200-5.
35. Bennett, C.L., et al., *The Research on Adverse Drug Events and Reports (RADAR) project.* Jama, 2005. **293**(17): p. 2131-40.
36. Gautier, S., et al., *The cost of adverse drug reactions.* Expert Opin Pharmacother, 2003. **4**(3): p. 319-26.
37. Lundkvist, J. and B. Jonsson, *Pharmacoeconomics of adverse drug reactions.* Fundam Clin Pharmacol, 2004. **18**(3): p. 275-80.
38. Pirmohamed, M., et al., *Adverse drug reactions.* Bmj, 1998. **316**(7140): p. 1295-8.
39. Pirmohamed, M. and B.K. Park, *Genetic susceptibility to adverse drug reactions.* Trends Pharmacol Sci, 2001. **22**(6): p. 298-305.
40. Goldstein, D.B., *Pharmacogenetics in the laboratory and the clinic.* N Engl J Med, 2003. **348**(6): p. 553-6.
41. Phillips, K.A., et al., *Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review.* Jama, 2001. **286**(18): p. 2270-9.
42. Shi, M.M., *Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies.* Clin Chem, 2001. **47**(2): p. 164-72.
43. Agrawal, M., et al., *Tyrosine kinase inhibitors: the first decade.* Curr Hematol Malig Rep. **5**(2): p. 70-80.
44. Cressey, T.R. and M. Lallemant, *Pharmacogenetics of antiretroviral drugs for the treatment of HIV-infected patients: an update.* Infect Genet Evol, 2007. **7**(2): p. 333-42.
45. Fontas, E., et al., *Lipid profiles in HIV-infected patients receiving combination antiretroviral therapy: are different antiretroviral drugs associated with different lipid profiles?* J Infect Dis, 2004. **189**(6): p. 1056-74.
46. Domingos, H., et al., *Metabolic effects associated to the highly active antiretroviral therapy (HAART) in AIDS patients.* Braz J Infect Dis, 2009. **13**(2): p. 130-6.

47. Carr, A., *Cardiovascular risk factors in HIV-infected patients.* J Acquir Immune Defic Syndr, 2003. **34 Suppl 1**: p. S73-8.
48. Calza, L., R. Manfredi, and F. Chiodo, *Dyslipidaemia associated with antiretroviral therapy in HIV-infected patients.* J Antimicrob Chemother, 2004. **53**(1): p. 10-4.
49. Barbaro, G., *Highly active antiretroviral therapy and the cardiovascular system: the heart of the matter.* Pharmacology, 2003. **69**(4): p. 177-9.
50. da Silva, E.F. and G. Barbaro, *New options in the treatment of lipid disorders in HIV-infected patients.* Open AIDS J, 2009. **3**: p. 31-7.
51. Tarr, P.E., et al., *Modeling the influence of APOC3, APOE, and TNF polymorphisms on the risk of antiretroviral therapy-associated lipid disorders.* J Infect Dis, 2005. **191**(9): p. 1419-26.
52. Arnedo, M., et al., *Contribution of 20 single nucleotide polymorphisms of 13 genes to dyslipidemia associated with antiretroviral therapy.* Pharmacogenet Genomics, 2007. **17**(9): p. 755-64.
53. Fellay, J., et al., *Response to antiretroviral treatment in HIV-1-infected individuals with allelic variants of the multidrug resistance transporter 1: a pharmacogenetics study.* Lancet, 2002. **359**(9300): p. 30-6.
54. Rotger, M., et al., *Contribution of genome-wide significant single-nucleotide polymorphisms and antiretroviral therapy to dyslipidemia in HIV-infected individuals: a longitudinal study.* Circ Cardiovasc Genet, 2009. **2**(6): p. 621-8.
55. Wahlund, S., *Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet.* Hereditas 1928. **11**: p. 65-106.
56. Azevedo, L., et al., *Epistatic interactions: how strong in disease and evolution?* Trends Genet, 2006. **22**(11): p. 581-5.
57. Cordell, H.J., *Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.* Hum Mol Genet, 2002. **11**(20): p. 2463-8.
58. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *International Joint Conference on Artificial Intelligence*. 1995.
59. Linden, A., *Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis.* J Eval Clin Pract, 2006. **12**(2): p. 132-9.
60. Lasko, T.A., et al., *The use of receiver operating characteristic curves in biomedical informatics.* J Biomed Inform, 2005. **38**(5): p. 404-15.
61. *Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III).* Jama, 2001. **285**(19): p. 2486-97.
62. Dube, M.P., et al., *Guidelines for the evaluation and management of dyslipidemia in human immunodeficiency virus (HIV)-infected adults receiving antiretroviral therapy: recommendations of the HIV Medical Association of the Infectious Disease Society of America and the Adult AIDS Clinical Trials Group.* Clin Infect Dis, 2003. **37**(5): p. 613-27.
63. Hastie, T., R. Tibshirani, and J. Friedman, *The Element of Statistical Learning: Data Mining, Inference and Prediction*. 2001, Springer.
64. Levy, D., et al., *Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study.* Hypertension, 2000. **36**(4): p. 477-83.
65. Tibshirani, R., *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 1996. **58**(1): p. 267-288.

66. Wee, L.J., T.W. Tan, and S. Ranganathan, *SVM-based prediction of caspase substrate cleavage sites.* BMC Bioinformatics, 2006. **7 Suppl 5**: p. S14.
67. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
68. Ehret, G.B., *Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension.* Curr Hypertens Rep. **12**(1): p. 17-25.
69. Newton-Cheh, C., et al., *Genome-wide association study identifies eight loci associated with blood pressure.* Nat Genet, 2009.
70. Todd, J.A., et al., *Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes.* Nat Genet, 2007. **39**(7): p. 857-64.
71. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.
72. McCarthy, M.I. and J.N. Hirschhorn, *Genome-wide association studies: potential next steps on a genetic journey.* Hum Mol Genet, 2008. **17**(R2): p. R156-65.
73. Thomas, G., et al., *Multiple loci identified in a genome-wide association study of prostate cancer.* Nat Genet, 2008. **40**(3): p. 310-5.
74. Easton, D.F., et al., *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature, 2007. **447**(7148): p. 1087-93.
75. Moffatt, M.F., et al., *Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma.* Nature, 2007. **448**(7152): p. 470-3.
76. Samani, N.J., et al., *Genomewide association analysis of coronary artery disease.* N Engl J Med, 2007. **357**(5): p. 443-53.
77. Kathiresan, S., et al., *Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans.* Nat Genet, 2008. **40**(2): p. 189-97.
78. Aulchenko, Y.S., et al., *Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts.* Nat Genet, 2009. **41**(1): p. 47-55.
79. Kathiresan, S., et al., *Common variants at 30 loci contribute to polygenic dyslipidemia.* Nat Genet, 2009. **41**(1): p. 56-65.
80. Weedon, M.N., et al., *Genome-wide association analysis identifies 20 loci that influence adult height.* Nat Genet, 2008. **40**(5): p. 575-83.
81. Willer, C.J., et al., *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation.* Nat Genet, 2009. **41**(1): p. 25-34.
82. Scuteri, A., et al., *Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits.* PLoS Genet, 2007. **3**(7): p. e115.
83. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.
84. Sebastiani, P., et al., *Genome-wide association studies and the genetic dissection of complex traits.* Am J Hematol, 2009. **84**(8): p. 504-15.
85. Firmann, M., et al., *The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome.* BMC Cardiovasc Disord, 2008. **8**: p. 6.
86. Carafoli, E., *Calcium-mediated cellular signals: a story of failures.* Trends Biochem Sci, 2004. **29**(7): p. 371-9.
87. Carafoli, E., *The ambivalent nature of the calcium signal.* J Endocrinol Invest, 2004. **27**(6 Suppl): p. 134-6.
88. Johnson, T. and Z. Kutalik, *QUICKTEST.* 2008.
89. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.

90. Bacanu, S.A., B. Devlin, and K. Roeder, *The power of genomic control.* Am J Hum Genet, 2000. **66**(6): p. 1933-44.
91. Saxena, R., et al., *Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge.* Nat Genet. **42**(2): p. 142-8.
92. Ingelsson, E., et al., *Detailed physiologic characterization reveals diverse mechanisms for novel genetic Loci regulating glucose and insulin metabolism in humans.* Diabetes. **59**(5): p. 1266-75.
93. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet. **42**(2): p. 105-16.
94. Alberti, K.G., P. Zimmet, and J. Shaw, *The metabolic syndrome--a new worldwide definition.* Lancet, 2005. **366**(9491): p. 1059-62.
95. Schwartz, M.W. and D. Porte, Jr., *Diabetes, obesity, and the brain.* Science, 2005. **307**(5708): p. 375-9.
96. Pischon, T., et al., *General and abdominal adiposity and risk of death in Europe.* N Engl J Med, 2008. **359**(20): p. 2105-20.
97. Loos, R.J. and C. Bouchard, *Obesity--is it a genetic disorder?* J Intern Med, 2003. **254**(5): p. 401-25.
98. Mutch, D.M. and K. Clement, *Unraveling the genetics of human obesity.* PLoS Genet, 2006. **2**(12): p. e188.
99. Rankinen, T., et al., *The human obesity gene map: the 2005 update.* Obesity (Silver Spring), 2006. **14**(4): p. 529-644.
100. Thorleifsson, G., et al., *Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity.* Nat Genet, 2009. **41**(1): p. 18-24.
101. Lindgren, C.M., et al., *Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution.* PLoS Genet, 2009. **5**(6): p. e1000508.
102. Carretero, O.A. and S. Oparil, *Essential hypertension. Part I: definition and etiology.* Circulation, 2000. **101**(3): p. 329-35.
103. Messerli, F.H., B. Williams, and E. Ritz, *Essential hypertension.* Lancet, 2007. **370**(9587): p. 591-603.
104. Cavalli-Sforza, L.L., *The genetics of human populations.* Sci Am, 1974. **231**(3): p. 80-9.
105. Tomson, J. and G.Y. Lip, *Blood pressure demographics: nature or nurture ... ... genes or environment?* BMC Med, 2005. **3**: p. 3.
106. Levy, D., et al., *Genome-wide association study of blood pressure and hypertension.* Nat Genet, 2009.

Appendices

Appendix 1: *Polymorphisms significantly associated with serum lipids, diabetes mellitus type 2, and/or obesity in recently published genome-wide analyses.*

Appendix 2: *Genome-wide meta-analysis for serum calcium identifies significantly associated SNPs near the calcium-sensing receptor (CASR) gene.*

Appendix 3: *Impact of atrial natriuretic peptide gene variants on HDL-cholesterol and other metabolic syndrome components in overweight/obese people. (Not available on the online version)*

Appendix 4: *Association analyses of 249,796 individuals reveal eighteen new loci associated with body mass index.*

Polymorphisms significantly associated with serum lipids, diabetes mellitus type 2, and/or obesity in recently published genome-wide analyses.

Note: The rs numbers of the 10 DM2 SNPs that are confirmed in 3 genome-wide analyses that studied genome variation using SNPs are highlighted bold. All participants in these genome-wide analyses were of Northern European ancestry. While all SNPs are located in genes involved in lipid and/or glucose metabolism, none of the indicated SNPs necessarily represents the causal variant – they may, rather, be in LD with or regulate the putative causal, functionally relevant variant(s). Most cited studies used gene "chips" covering at least 300K or 500K SNPs - A study using an array with reduced genome coverage (Affymetrix 100K GeneChip) failed to identify any gene-lipid associations.

rs number	Gene or closest gene	Phenotype predominantly affected by variant allele(s)	Ref.	No. of SNPs that will be included in the analysis of the following endpoints:						Comments
				HDL-C	Non-HDL-C	TG	Glucose/DM2	Met Sy	Anthropometric changes	
rs780094	GCKR	TG	[1-3]			1		1		glucokinase regulatory protein: regulates glucokinase
rs429358 rs7412	APOE/C1/C4	LDL-C	[1-3]		3					Apolipoprotein E, APOE cluster
rs4420638***										*** rs4420638 is in strong linkage disequilibrium with rs429358 and rs7412 and will therefore not be genotyped.
rs10402271										rs708272 and rs7412 were interrogated in Lipogen 2 [4]
rs693	APOB	LDL-C, TG	[1-3]		3	3		3		rs693 = APOB C7673T
rs562338 rs754523										Only LDL-associated SNP that was not enriched in patients with coronary artery disease vs controls in [3]
rs1800775	CETP	HDL-C	[1, 2]	6				6		Cholesteryl Ester Transfer Protein
rs3764261 rs1864163 rs9989419 rs12596776 rs1566439										rs1800775 (-629C>A) and rs708272 (279G>A) were interrogated in Lipogen 2 [4]
rs708272		TG	[4]			1				
rs328	LPL	HDL-C, TG	[1-3]	3		3		3		Lipoprotein Lipase, LPL S447X
rs2197089 rs6586891										rs328 was interrogated in Lipogen 2 [4] Another SNP is: rs10503669, which is in perfect LD with rs328
rs1800588 rs4775041	LIPC	HDL-C, TG	[1-3]	2		2		2		Hepatic Lipase, LIPC C-514T
										Another SNP is rs261332, which is in strong LD with rs1800588
rs2266788 rs12286037 rs28927680 rs3135506	APOA5/A4/C3/A1 ZNF259, BUD13	HDL-C, TG	[2, 3]	4		4 3		4		rs2266788 = APOA5 T1259C rs662799 (APOA5 -1131T>C) and rs3135506 (APOA5 64C>G) were interrogated in Lipogen 2 [4]
rs2854117 rs45537037rs5128 rs662799			[4]			1				rs2854117 (APOC3 -482C>T), rs45537037 (APOC3 -455T>C), and rs5128 (APOC3 +3238C>G) were interrogated in Lipogen 2 [4]

rs11591147 rs11206510	PCSK9	LDL-C	[1, 3]		2				Another SNP: rs505151 was associated with LDL-C and stroke in non-genome-wide analyses [5-7]. LDL-C association not confirmed in [8] Proprotein convertase subtilisin/kexin type 9 rs11591147 = E670G = 23968A>G
rs646776	CELSR2, PSRC1, SORT1	LDL-C	[1]		1				2 additional SNPs contribute to LDL-C; they are in strong LD with rs646776: rs599838 [1] rs599839 [3]
rs16996148	NCAN, CILP2, PBX4	LDL-C, TG	[1, 3]		1	1		1	
rs4846914	GALNT2	HDL-C, TG	[1, 3]	1		1		1	Another SNP is rs2144300, which is in perfect LD with rs4846914
rs17145738	BCL7B, TBL2, MLXIPL	TG	[1, 3]			1		1	
rs17321515	TRIB1	TG	[1, 3]			1		1	
rs12130333 rs1748195	ANGPTL3, DOCK7, ATG4C	TG	[1, 3]			2		2	
rs12654264	HMGCR	LDL-C	[1]		1				
rs6511720	LDLR	LDL-C	[1, 3]		1				
rs3890182 rs4149268 rs4149313	ABCA1	HDL-C TG	[1, 3] [4]	2		 1		2	 rs4149313 (2962A>G) was interrogated in Lipogen 2 [4]
rs2156552 rs2000813	LIPG, ACAA2	HDL-C Non-HDL-C	[1, 3] [4]	1	 1			1	 rs2000813 (584C>T) was interrogated in Lipogen 2 [4]
rs2338104	MVK/MMAB	HDL-C	[3]	1				1	
rs8050136	FTO	DM2, obesity	[9-11]				1	1	1	risk allele=A, non-risk=C Fat mass and obesity associated, FTO: a gene of unknown function [9] rs8050136 associated with BMI in DM2 and controls, and with waist circumference in cases only. DM2 association of rs8050136 disappears when adjusted for BMI and waist circumference [11] rs9939609 used in for all analyses since all other SNPs highly correlated and because of 100% genotyping success rate. rs9939609 is highly correlated with 45 SNPs in the first FTO intron [9] Hapmap population frequency of risk allele was 0.45 in Europeans, 0.14 in Japanese and Chinese [9]
rs4430796	TCF2	DM2	[12]				1	1		No anthropometric information
rs7903146	TCF7L2	DM2	[2, 10, 11, 13]				1	1		risk allele= T, non-risk=C. Transcription factor 7-like 2 Variant with the highest odds ratio for DM2 (1.37), substantially higher than the other “DM2 variants”. A second SNP in strong LD with rs7903146 is: rs7901695 [11]

rs13266634	SLC30A8	DM2	[2, 10, 11, 13, 14]	1	1	risk allele=C, non-risk=T Solute carrier family 30, member 8 SLC30A8 R325W
rs1111875	IDE-KIF11-HHEX	DM2	[2, 10, 11, 13]	1	1	risk allele=C, non-risk=T Contained in an LD block containing insulin degrading enzyme (IDE), the homeodomain protein HHEX, and the kinesin-interacting factor KIF11 Additional risk alleles in this LD block are: rs5015480 rs7923837
rs9300039	EXT2-ALX4-LOC387761	DM2	[10, 13]	1	1	Contained in an LD block containing genes potentially involved in beta cell function or development rs9300039 is located in an intergenic region on chromosome 11; in zero LD with rs7480010 [10]; in LD with 58 SNPs [10] (risk allele=C, non-risk=A) Additional risk alleles in this LD block are: 3 SNPs located in introns of exostosin 2 (EXT2) (rs3740878, rs11037909 and rs1113132); rs7480010 (in/near a hypothetical gene LOC387761) [13]; NB: DM2 association of rs7480010 not replicated in [10]; DM2 association of the 3 intronic variants of EXT2 not replicated in [13]
rs10811661	CDKN2A/B	DM2	[2, 10, 11]	1	1	risk allele=T, non-risk=C Cyclin dependent kinase inhibitor 2A/B A second SNP in strong LD with rs 10811661 is: rs564398
rs4402960	IGF2BP2	DM2	[2, 10, 11] [2] for rs1470579	1	1	risk allele=T, non-risk=G insulin-like growth factor 2 mRNA-binding protein 2 A second SNP in strong LD with rs4402960 is: rs1470579
rs7754840	CDKAL1	DM2	[2, 10, 11, 14]		1	risk allele=C, non-risk=G CDK5 regulatory subunit-associated protein 1-like 1 associated with waist circumference in DM2 cases but not controls [11] Two additional SNPs in strong LD with rs7754840 are: rs10946398, rs7756992
Rs5219	KCNJ11	DM2	[2, 10, 11, 15]	1	1	risk allele=T, non-risk=C (=KCNJ11 E23K) Potassium Channel, Inwardly Rectifying, Subfamily J, Member 11 A second SNP in strong LD with rs5219 is rs5215 [2]
rs1801282	PPARG	DM2	[2, 10, 11, 16]	1	1	risk allele=C, non-risk=G = PPARG P12A Peroxisome proliferative activated receptor, gamma, PPARG [16]: not a genome-wide analysis

rs10010131 rs6446482	WFS1	DM2	[17]				2	2	not a genome-wide analysis Wolframin, mutations of which are associated with Wolfram syndrome, characterized by DM and diabetes insipidus P-values for DM2 association of the two SNPs are, respectively, 1.4 x10-7 and 3.4 x10-7
Total Number of SNPs genotyped for each metabolic endpoint				**20**	**13**	**25**	**12**	**42**	**1**

References

1. Kathiresan, S., et al., *Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans.* Nat Genet, 2008. **40**(2): p. 189-97.
2. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.
3. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* Nat Genet, 2008. **40**(2): p. 161-9.
4. Arnedo, M., et al., *Contribution of 20 single nucleotide polymorphisms of 13 genes to dyslipidemia associated with antiretroviral therapy.* Pharmacogenet Genomics, 2007. **17**(9): p. 755-64.
5. Abboud, S., et al., *Proprotein convertase subtilisin/kexin type 9 (PCSK9) gene is a risk factor of large-vessel atherosclerosis stroke.* PLoS One, 2007. **2**(10): p. e1043.
6. Chen, S.N., et al., *A common PCSK9 haplotype, encompassing the E670G coding single nucleotide polymorphism, is a novel genetic marker for plasma low-density lipoprotein cholesterol levels and severity of coronary atherosclerosis.* J Am Coll Cardiol, 2005. **45**(10): p. 1611-9.
7. Evans, D. and F.U. Beil, *The E670G SNP in the PCSK9 gene is associated with polygenic hypercholesterolemia in men but not in women.* BMC Med Genet, 2006. **7**: p. 66.
8. Kotowski, I.K., et al., *A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol.* Am J Hum Genet, 2006. **78**(3): p. 410-22.
9. Frayling, T.M., et al., *A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity.* Science, 2007. **316**(5826): p. 889-94.
10. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-5.
11. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.
12. Gudmundsson, J., et al., *Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes.* Nat Genet, 2007. **39**(8): p. 977-83.
13. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-5.
14. Steinthorsdottir, V., et al., *A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.* Nat Genet, 2007. **39**(6): p. 770-5.
15. Florez, J.C., et al., *Type 2 diabetes-associated missense polymorphisms KCNJ11 E23K and ABCC8 A1369S influence progression to diabetes and response to interventions in the Diabetes Prevention Program.* Diabetes, 2007. **56**(2): p. 531-6.
16. Florez, J.C., et al., *Effects of the type 2 diabetes-associated PPARG P12A polymorphism on progression to diabetes and response to troglitazone.* J Clin Endocrinol Metab, 2007. **92**(4): p. 1502-9.
17. Sandhu, M.S., et al., *Common variants in WFS1 confer risk of type 2 diabetes.* Nat Genet, 2007. **39**(8): p. 951-3.

OPEN ACCESS Freely available online

PLoS GENETICS

Genome-Wide Meta-Analysis for Serum Calcium Identifies Significantly Associated SNPs near the Calcium-Sensing Receptor (*CASR*) Gene

Karen Kapur[1,2], Toby Johnson[1,2,3], Noam D. Beckmann[1], Joban Sehmi[4], Toshiko Tanaka[5], Zoltán Kutalik[1,2], Unnur Styrkarsdottir[6], Weihua Zhang[7], Diana Marek[1,2], Daniel F. Gudbjartsson[6], Yuri Milaneschi[5], Hilma Holm[6], Angelo Dilorio[8], Dawn Waterworth[9], Yun Li[10], Andrew B. Singleton[11], Unnur S. Bjornsdottir[12], Gunnar Sigurdsson[13,14], Dena G. Hernandez[11], Ranil DeSilva[4], Paul Elliott[15], Gudmundur I. Eyjolfsson[12], Jack M. Guralnik[16], James Scott[4], Unnur Thorsteinsdottir[6,13], Stefania Bandinelli[17], John Chambers[7], Kari Stefansson[6,13], Gérard Waeber[18], Luigi Ferrucci[5], Jaspal S. Kooner[4], Vincent Mooser[9], Peter Vollenweider[18], Jacques S. Beckmann[1,19], Murielle Bochud[3☯], Sven Bergmann[1,2*☯]

1 Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** Institute of Social and Preventive Medicine, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland, **4** National Heart and Lung Institute, Imperial College London, London, United Kingdom, **5** Clinical Research Branch, National Institute on Aging, Baltimore, Maryland, United States of America, **6** deCODE Genetics, Reykjavik, Iceland, **7** Department of Epidemiology and Public Health, Imperial College London, London, United Kingdom, **8** Department of Medicine and Sciences of Aging, Laboratory of Clinical Epidemiology, University G. d'Annunzio, Chieti, Italy, **9** Division of Genetics, GlaxoSmithKline, King of Prussia, Pennsylvania, United States of America, **10** Department of Genetics and Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, United States of America, **11** Laboratory of Neurogenetics, National Institute on Aging, Bethesda, Maryland, United States of America, **12** The Laboratory in Mjodd, RAM, Reykjavik, Iceland, **13** Faculty of Medicine, University of Iceland, Reykjavik, Iceland, **14** Department of Endocrinology and Metabolism, University Hospital, Reykjavik, Iceland, **15** Department of Epidemiology and Biostatistics, Medical Research Council–Health Protection Agency Centre for Environment and Health, Imperial College London, London, United Kingdom, **16** Laboratory of Epidemiology, Demography, and Biometry, National Institute on Aging, Bethesda, Maryland, United States of America, **17** Geriatric Unit, Azienda Sanitaria Firenze, Florence, Italy, **18** Department of Medicine, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland, **19** Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

Abstract

Calcium has a pivotal role in biological functions, and serum calcium levels have been associated with numerous disorders of bone and mineral metabolism, as well as with cardiovascular mortality. Here we report results from a genome-wide association study of serum calcium, integrating data from four independent cohorts including a total of 12,865 individuals of European and Indian Asian descent. Our meta-analysis shows that serum calcium is associated with SNPs in or near the calcium-sensing receptor (*CASR*) gene on 3q13. The top hit with a p-value of 6.3×10^{-37} is rs1801725, a missense variant, explaining 1.26% of the variance in serum calcium. This SNP had the strongest association in individuals of European descent, while for individuals of Indian Asian descent the top hit was rs17251221 ($p=1.1\times10^{-21}$), a SNP in strong linkage disequilibrium with rs1801725. The strongest locus in *CASR* was shown to replicate in an independent Icelandic cohort of 4,126 individuals ($p=1.02\times10^{-4)}$. This genome-wide meta-analysis shows that common *CASR* variants modulate serum calcium levels in the adult general population, which confirms previous results in some candidate gene studies of the *CASR* locus. This study highlights the key role of *CASR* in calcium regulation.

Citation: Kapur K, Johnson T, Beckmann ND, Sehmi J, Tanaka T, et al. (2010) Genome-Wide Meta-Analysis for Serum Calcium Identifies Significantly Associated SNPs near the Calcium-Sensing Receptor (*CASR*) Gene. PLoS Genet 6(7): e1001035. doi:10.1371/journal.pgen.1001035

Editor: Gonçalo R. Abecasis, University of Michigan, United States of America

Received November 9, 2009; **Accepted** June 17, 2010; **Published** July 22, 2010

Funding: The CoLaus study is supported by GlaxoSmithKline, by the Faculty of Biology and Medicine of the University of Lausanne, Switzerland, and by a grant from the Swiss National Science Foundation: 33CSCO-122661. JS Beckman is supported by a grant form the Swiss National Foundation (310000-112552). M Bochud is supported by grants from the Swiss National Foundation (PROSPER 3200BO-111362/1, 3233BO-111361/1) and by the Swiss School of Public Health Plus (SSPH+). S Bergmann is grateful for financial support from the Giorgi-Cavaglieri Foundation, the Swiss National Science Foundation (Grant #3100AO-116323/1), the Swiss Institute of Bioinformatics and the European Framework Project 6 (through the AnEuploidy and EuroDia projects). The BLSA was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. A portion of that support was through a R&D contract with Medstar Research Institute. The InCHIANTI study baseline (1998–2000) was supported as a "targeted project" (ICS110.1/RF97.71) by the Italian Ministry of Health and in part by the US National Institute on Aging (Contracts: 263 MD 9164 and 263 MD 821336). Y Li is partially supported by DK078150-03 and DK056350 (to the University of North Carolina Nutrition Obesity Research Center). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Vincent Mooser and Dawn Waterworth are full-time employees of GlaxoSmithKline, a pharmaceutical company. Peter Vollenweider and Gerard Waeber received financial support from GlaxoSmithKline to build the CoLaus study. Daniel F Gudbjarsson, Unnur Styrkarsdottir, Hilma Holm, Unnur Thorsteinsdottir and Kari Stefansson are employees of deCODE genetics, a biotechnology company.

* E-mail: Sven.Bergmann@unil.ch

☯ These authors contributed equally to this work.

Introduction

Calcium is the most abundant mineral in the human body contributing approximately one kilogram to the average adult human body mass. Whereas 99% of calcium is stored in the skeleton and teeth, the remaining 1% circulates in the bloodstream and is involved in many physiological processes including its function as a universal cellular signaling molecule [1–2]. Calcium plays a key role in membrane potential, which is important for muscle contraction, heart rate regulation and generation of nerve impulses. Calcium also influences bone metabolism, ion transport and many other cellular processes [3]. Approximately 2/5 of calcium in the extracellular fluid is found in blood serum. The level of serum calcium is under tight hormonal control with a normal concentration of 2.15–2.55 mmol/L. Serum calcium is under strong genetic control, with twin studies showing that the variance in total calcium due to genetic effects is between 50% and 78% [4–5].

While skeletal calcium is important in numerous clinical disorders, in particular bone and mineral disorders, the clinical role of serum calcium is less clear. Several [6–7] (but not all [8]) studies indicated that elevated serum calcium levels are associated with an increased risk of cardiovascular disease. Patients with hyperparathyroidism, who suffer from chronic hypercalcemia, have a high prevalence of hypertension and increased cardiovascular mortality [9–11]. However, the mechanisms underlying the putative association of serum calcium with increased cardiovascular morbidity and mortality remain unclear.

Rare monogenic forms of hypo- or hypercalcemia have been described, including disorders involving the calcium-sensing receptor (*CASR*, locus 3q13) gene. Heterozygous and homozygous *CASR* mutations that inactivate CASR are responsible, respectively, for familial hypocalciuric hypercalcemia, type 1 (also known as familial benign hypercalcemia) (OMIM #145980) [12–13] and neonatal severe hyperparathyroidism (OMIM #239200) [13]. On the other hand, mutations that result in CASR activation lead to autosomal dominant hypocalcemia (OMIM #146200) [14]. Mutations in many other genes have also been found to lead to disturbances of serum calcium levels (Table 1).

In the present study, we report results obtained from meta-analysis of genome-wide associations of serum calcium levels from four cohorts with a total of 12,865 participants. We first describe the design of the study and its main finding, that variants in *CASR* give rise to the strongest signals associated with serum calcium levels in both European and Indian Asian populations. Our results confirm previous studies showing that mutations in *CASR* are associated with serum calcium levels in young healthy women [15–16] and extend this observation to men and women across a large spectrum of age. We show that *CASR* is a key player in the genetic regulation of serum calcium in men and women from the general adult population.

Results

We performed a meta-analysis for genome-wide associations of serum calcium, determined by subtracting the estimated amount of calcium bound to albumin from the total serum calcium, to infer the amount of ionized calcium (see Materials and Methods). Our study included four cohorts: (i) 5404 European individuals from the Cohorte Lausanne (CoLaus) [17–18], (ii) 5548 European and Indian Asian individuals from the London Life Sciences Population (LOLIPOP) Study from West London UK [19–20], (iii) 1196 European individuals from the InCHIANTI Study (Tuscany, Italy) [21], and (iv) 717 individuals of European descent from the Baltimore Longitudinal Study of Aging (BLSA) study based in the Baltimore-Washington DC area [22], totaling 12,865 participants (see Table 2 for more detailed characteristics of each cohort).

Genome-wide association scans were performed first independently for each cohort using linear regression and then the effect sizes from each cohort were meta-analyzed (see Materials and Methods). Due to the possibility of population substructure obscuring effects of genetic variants, meta-analysis was performed separately for (i) combined European and Indian Asian cohorts (N = 12,865) and restricted to cohorts of (ii) European (N = 8,919), and (iii) Indian Asian descent (N = 3,947). The meta-analyses yielded 100 SNPs from the combined cohorts, 70 SNPs when restricting to European cohorts and 22 SNPs restricting to Indian Asian cohorts that exceeded the genome-wide significance threshold of 5×10^{-8} (Figure 1A–1C) (the full list is provided in Table S2A, S2B, S2C). All SNPs reaching statistical significance clustered around the *CASR* locus at 3q13. The most significant SNP in the (i) combined and (ii) European meta-analyses was rs1801725 ($p = 6.29\times10^{-37}$, $p = 2.58\times10^{-18}$, respectively) and in the (iii) Indian Asian meta-analysis was rs17251221 ($p = 1.07\times10^{-21}$). These two SNPs are less than 11 kb apart and are in high linkage disequilibrium with each other ($r^2 = 0.946$, 0.494, 1.0, 1.0 in HapMap CEU, CHB, JPT, YRI, respectively), and therefore most likely derive from the same association signal. We find that rs1801725 explains 1.26% of the variance in serum calcium, with the effect sizes and standard errors of the serum calcium increasing *T* allele in individual cohorts shown in Figure 2 and Table S3. According to our additive model, each rs1801725 *T* allele increases $\log_{10}$ serum calcium (in units mmol/L) by 3.61×10^{-3}, equivalent to a multiplicative effect of 1.008 on serum calcium (see also Table S2). At an average serum calcium level of 2.25 mmol/L, each rs1801725 *T* allele yields an increase of 0.01874 mmol/L, or 21% of one standard deviation of serum calcium levels in a normal population. The regional pattern of association of SNPs around the *CASR* locus, and their linkage disequilibrium with rs1801725, are shown in Figure 3. Of note, rs1042636, which has been associated with decreased serum calcium [23], also achieved genome-wide significance with the G minor allele associated with decreased serum calcuim ($p = 4.96\times10^{-9}$). However, conditional on the rs1801725 locus, located 12 bps upstream, the rs1042636 p-value became 3.32×10^{-4}, indicating that the two loci share contributions to serum calcium levels.

To confirm the rs1801725 signal, we analyzed the association pattern with serum calcium in a separate cohort. We used a subset of 4,126 Icelandic individuals from the deCODE study [24–25] with serum calcium measurements. We found the rs1801725 *T* allele to be strongly associated with increased serum calcium ($p = 1.02\times10^{-4}$), replicating the key meta-analysis result.

While only the *CASR* locus reached nominal genome-wide significance for association with serum calcium, the top regions with $p<10^{-5}$ are shown in Table 3. These SNPs cover 12 regions, the significance of which is displayed across cohorts in Figure S3. There were no SNPs in other candidate genes (which have previously been shown to be involved in disorders associated with disturbed serum calcium levels) that were associated with serum calcium at genome-wide significance. The most significant SNPs within 500 kb of the gene transcripts are shown in Table 1. Considering the set of 18,611 distinct SNPs mapping to the set of serum calcium candidate genes excluding *CASR*, we find no significant association (at significance level 0.05 and applying the

Author Summary

Calcium levels in blood serum play an important role in many biological processes. The regulation of serum calcium is under strong genetic control. This study describes the first meta-analysis of a genome-wide association study from four cohorts totaling 12,865 participants of European and Indian Asian descent. Confirming previous results in some candidate gene studies, we find that common polymorphisms at the calcium-sensing receptor (*CASR*) gene locus are associated with serum calcium concentrations. We show that *CASR* variants give rise to the strongest signals associated with serum calcium levels in both European and Indian Asian populations, while no other locus reaches genome-wide significance. Our results show that *CASR* is a key player in genetic regulation of serum calcium in the adult general population.

Bonferroni correction for multiple testing, giving a cut-off p-value of 2.69×10^{-6}, see also Figure S4). Indeed, fixing the sample size and genome-wide significance threshold our study is well-powered ($\geq$0.80) to detect SNPs explaining at least 0.31% of the variance. Therefore the common SNPs within the candidate genes (excluding CASR) likely play at best a small role in serum calcium regulation.

We analyzed the association of the top SNP with several calcium-related outcomes (coronary heart disease, myocardial infarction, hypertension, stroke, osteoarthritis, osteoporosis and kidney stones). The number of cases and controls for each outcome and each cohort is given in Table S4. Logistic regression including age and pseudosex (see Materials and Methods) as covariates did not find any significant association between rs1801725 and the calcium-related outcomes, after correcting for multiple testing (effect sizes and standard errors for the *T* allele are listed in Table S5). Power calculations show that given the sample sizes for the clinical traits above, our study has good power ($\geq$0.80) to detect odds ratios of 1.20, 1.13, 1.77, 1.27, 1.27, 1.24 and 1.75, respectively. As the smallest p-values from calcium-related traits were for osteoarthritis and osteoporosis (bonferroni-corrected p = 0.21, 0.44, respectively), we further investigated bone density traits. None of deCODE hip bone mineral density or spine bone mineral density (N = 6657 and 6838, respectively) nor In-CHIANTI total bone density, trabecular bone density, cortical bone density, cortical bone thickness or cortical bone area (N = 1196) bonferroni-adjusted p-values for eight traits were significant.

Discussion

This genome-wide scan of 12,865 individuals revealed *CASR* as the most significant (and only genome-wide significant) locus influencing serum calcium levels. Specifically, we found evidence for a strong association of SNPs located in the *CASR* locus with serum calcium levels in both Europeans and Indian Asians. The strongest locus in *CASR* was further shown to replicate in an independent Icelandic cohort of 4,126 individuals.

The top signal (rs1801725, *2956G>T*) explains 1.26% of the variance in serum calcium. Indeed, this is similar to results from other GWAS of human height [26–29], body mass index [30–31], serum urate [32–34] and serum lipid concentrations [34–36], for which the genome-wide significant loci uncovered thus far collectively explain only a small fraction of the phenotypic variance (usually at least one order of magnitude less than the total additive genetic variance estimated from heritability studies [37–38]). The rs1801725 *T* allele (A986S) was associated with higher serum calcium, consistent with previous findings (see Table S6). The rs1801725 polymorphism (with *T* allele frequencies of 16.76%, 19.98% in European and Indian Asian cohorts, respectively) affects serum calcium levels of a substantial proportion of the population.

The rs1801725 polymorphism encodes a missense variant in exon 7 of the *CASR* gene, which leads to a non-conservative amino-acid change (serine substitution for alanine-986, A986S corresponding to nucleotides *2956G>T*) in the cytoplasmic tail of CASR. *In vitro* studies showed that mutations within the C-terminal tail may influence several aspects of CASR function, such as signal transduction, intracellular trafficking and cell surface expression [39–41]. However, PolyPhen predicts rs1801725 to be a benign substitution. It is presently unclear whether this substitution gives rise to functional variants, as functional studies have yielded conflicting results [42–43]. Deep sequencing of this region may help identifying the causal variants. While it is still not possible to infer a direct causal role, it is of interest to note that the SNP gives rise to an amino acid change in the C-terminal tail of CASR, a domain which plays a key role in the receptor function and may potentially influence intracellular trafficking following CASR activation by extracellular calcium.

Several studies have reported associations of A986S and nearby *CASR* mutations with various phenotypes. The A986S *CASR* polymorphism has been associated with variations in circulating calcium levels in healthy adults in some studies [15,23,44–45], but not in others [46–47]. The fact that the latter studies were underpowered (sample size ranging from 148 to 1252) to detect a small effect size likely explains these inconsistent results. The rs1042636 (R990G) polymorphism has been associated with the magnitude of parathyroid hormone (PTH) secretion in patients with primary hyperparathyroidism [48], and preliminary results suggest that it could influence response to cinacalcet, a calcimimetic used to treat secondary hyperparathyroidism in patients with end-stage renal disease [49]. In a meta-analysis, 986S was associated with a 49% increased risk (P = 0.002) of primary hyperparathyroidism [47,50–51]. Among patients with primary hyperparathyroidism, the AGQ haplotype (i.e. 986A, 990G, 1011Q, which is associated with lower serum calcium levels and hypercalciuria [52]) was associated with increased risk, and the SRQ haplotype with decreased risk, of kidney stones [50].

CASR has been previously considered as a candidate gene for osteoporosis [53] and coronary heart disease as well as increased total and cardiovascular mortality [54]. In our meta-analysis, we found no significant association of rs1801725 with these calcium-related phenotypes. A recent meta-analysis focusing on effects of candidate genes on osteoporosis also reports negative results for *CASR* [55]. Furthermore, results on the association of elevated serum calcium with increased cardiovascular risk in the general population are controversial [6–8]. It is therefore not clear to what extent serum calcium might predict cardiovascular risk. The SNPs identified in this meta-analysis could serve as genetic instruments in future studies, such as Mendelian randomization analysis in longitudinal cohorts, to further investigate the causal effect of serum calcium on osteoporosis and on cardiovascular disease risk (see Table S5 for rs1801725 effects and standard errors).

Our meta-analysis suffers from some limitations. First, we used corrected serum calcium and not directly measured ionized serum calcium. The correlation between corrected serum calcium and

Table 1. Serum calcium candidate genes.

Gene	Gene region	Disease	OMIM	# SNPs*	Top SNP†	Top SNP p-value†
AIRE	chr21: 44,530,191-44,542,530	Autoimmune polyendocrine syndrome, type I	240300	870	rs2838473, rs13052277, rs717177	2.48E-04, 2.78E-05, 4.55E-03
ALPL	chr1: 21,581,175-21,650,208	Hypophosphatasia, infantile	241500	786	rs6426723, rs1256348, rs4654973	6.88E-03, 1.14E-03, 7.30E-04
BSND	chr1: 55,176,638-55,186,485	Bartter syndrome, type 4	602522	1007	rs17111592, rs11584093, rs6588528	4.61E-03, 8.40E-05, 1.70E-02
CASR	chr3:123,385,220-123,488,032	Familial hypocalciuric hypercalcemia, type 1; neonatal severe hyperparathyroidism; autosomal dominant hypocalcemia	145980; 239200; 146200; 241200	921	rs1801725, rs1801725, rs17251221	6.29E-37, 2.58E-18, 1.07E-21
CDKN1B	chr12: 12,761,576-12,766,569	Bartter syndrome, type 4	131100	855	rs3825271, rs888200, rs11055225	7.05E-04, 3.56E-03, 8.66E-04
CLCNKA	chr1: 16,093,672-16,105,850	Bartter syndrome, type 4	602522	545	rs12405694, rs16852052, rs6661012	1.34E-02, 2.51E-02, 1.00E-02
CLCNKB	chr1: 16,115,658-16,128,782	Bartter syndrome, type 3; Bartter syndrome, type 4	607364; 602522	520	rs12405694, rs16852052, rs6661012	1.34E-02, 2.51E-02, 1.00E-02
CLDN16	chr3: 191,588,543-191,611,035	Hypomagnesemia 3, renal	248250	1381	rs11714779, rs11714779, rs9682599	3.15E-04, 2.86E-04, 3.18E-02
CYP27B1	chr12: 56,442,384-56,447,145	Vitamin D-dependant rickets type I	264700	567	rs11172284, rs810204, rs715930	2.18E-03, 2.36E-03, 3.26E-02
GATA3	chr10: 8,136,673-8,157,170	Hypoparathyroidism; sensorineural deafness; renal disease	146255	1492	rs11812109, rs12359361, rs2765399	1.30E-03, 2.86E-03, 4.35E-03
GCM2	chr6: 10,981,450-10,990,084	Familial hyperparathyroidism	146200	908	rs16870899, rs16870899, rs6457160	4.88E-03, 4.59E-03, 3.36E-03
GNAS	chr20: 56,900,130-56,919,640	Pseudohypoparathyroidism, type IA; Pseudohypoparathyroidism, type IB	103580; 603233	949	rs2145477, rs911297, rs6015375	2.38E-04, 3.95E-03, 1.12E-02
HRPT2	chr1: 189,822,81-189,952,713	Hyperparathyroidism (familial isolated hyperparathyroidism); parathyroid carcinoma	145000; 608266	836	rs10737627, rs913478, rs2887613	1.15E-02, 6.50E-03, 4.71E-02
KCNJ1	chr11: 128,213,12-128,242,478	Bartter syndrome, antenatal, type 2	241200	1059	rs948215, rs3897566, rs7116606	1.08E-03, 1.50E-03, 3.46E-05
MEN1	chr11: 64,327,572-64,335,342	Hyperparathyroidism (familial isolated hyperparathyroidism); multiple endocrine neoplasia, type I	145000; 131100	498	rs7947143, rs7947143, rs11820322	1.49E-02, 6.83E-03, 2.92E-03
PHEX	chrX: 21,810,216-22,025,985	Hypophosphatemic rickets, X-linked dominant	307800	NA	NA	NA
PTH	chr11: 13,470,178-13,474,143	Familial hyperparathyroidism	146200	1292	rs10832087, rs10500780, rs1502242	9.61E-04, 1.71E-03, 1.96E-03
PTH1R	chr3: 46,894,240-46,920,291	Jansen's metaphyseal chondrodysplasia	156400	511	rs1402151, rs883739, rs6442037	1.03E-02, 7.54E-03, 2.18E-02
RET	chr10: 42,892,533-42,944,955	Multiple endocrine neoplasia, type I	131100	856	rs3026762, rs3026762, rs12265792	2.10E-04, 1.90E-04, 4.10E-02
SLC12A1	chr15: 46,285,790-46,383,568	Bartter syndrome, antenatal, type I	601678	925	rs1025759, rs596942, rs919129	2.25E-03, 1.29E-02, 2.30E-04
SLC4A1	chr17: 39,682,566-39,700,993	Renal tubular acidosis, distal, autosomal dominant	179800	494	rs12602991, rs12602991, rs708384	2.18E-03, 7.20E-03, 4.05E-03
TBCE	chr1: 231,856,81-231,938,321	Hypoparathyroidism-retardation-dysmorphism syndrome	241410	591	rs12133603, rs12133603, rs291353	2.34E-03, 2.18E-03, 5.38E-03
TRPM6	chr9: 74,566,965-74,732,564	Hypomagnesemia with secondary hypocalcemia	602014	1123	rs877809, rs877809, rs12550903	9.31E-04, 1.20E-03, 6.14E-03
VDR	chr12: 46,521,589-46,585,081	Vitamin D-resistant rickets type II; vitamin D-dependent rickets, type II	277440; 259700	1066	rs1859441, rs1859441, rs11168354	1.22E-03, 9.59E-04, 3.34E-04

Genes which have been shown to lead to disturbances of serum calcium levels. For each gene, we report the top SNP for the meta-analyses of all cohorts, European cohorts only and Indian Asian cohorts only.
*500 kb upstream and downstream the gene region.
†Combined, European, and Indian Asian.
doi:10.1371/journal.pgen.1001035.t001

Table 2. Characteristics of participants, by study.

	CoLaus	LOLIPOP European Whites	LOLIPOP Indian Asians	InCHIANTI	BLSA	deCODE
Sample size	5404	1601	3947	1196	717	4126
Gender (males/females)	2542/2862	1397/204	3832/115	533/663	390/327	1313/2813
Age (years)*	53.43 (34.9,75.4)	54.5 (22.6,75.0)	50.7 (35.0,74.9)	68.22 (21,102)	70.4 (22,98)	60.1 (7,103)
Pre/Post menopause	1210/1652	100/104	56/59	79/584	38/289	872/1836 (105 with pre and post measurements)
Serum calcium (mmol/L)*	2.29 (0.094)	2.41 (0.12)	2.37 (0.11)	2.36 (0.10)	2.31 (0.11)	2.38 (0.14)
Corrected serum Calcium (mmol/L)*	2.18 (0.09)	2.31 (0.09)	2.29 (0.09)	2.30 (0.09)	2.30 (0.1)	2.31 (0.14)
Serum albumin [g/L]*	44.2 (2.5)	43.7 (2.9)	43.4 (2.9)	42.3 (3.1)	40.4 (3.5)	42.9 (3.9)

Characteristics are shown for CoLaus, LOLIPOP European, LOLIPOP Indian Asian, InCHIANTI, BLSA and deCODE. Corrected serum calcium, designed to estimate the amount of biologically active serum calcium, is defined as Ca_corrected = total serum calcium [mmol/L] + (40 - albumin [g/L])/40.
*Values represent mean (range) or mean (sd).
doi:10.1371/journal.pgen.1001035.t002

ionized serum calcium varies between 0.66 and 0.87 [56–58]. We can hypothesize that the association of ionized serum calcium with *CASR* variants would be stronger than the one with corrected serum calcium because ionized calcium is the form physiologically active on CASR. Second, data on serum phosphate, PTH or vitamin D are not available, so that we cannot explore further these relationships. Third, sample sizes for calcium-related clinical traits were limited, many clinical traits in CoLaus were self-reported instead of clinically diagnosed, and we incur a multiple testing penalty due to the number of clinical traits posited to be associated with serum calcium. However, the major strengths of the study are the hypothesis-free nature of GWAS studies, the large sample meta-analysis and the inclusion of multiple populations.

Materials and Methods

Cohorts

CoLaus is a population-based sample from Lausanne, Switzerland, consisting of 5435 individuals between 35 and 75 years old (after QC) of which a subset of 5404 had available serum calcium measurements. The study design and protocols have been described previously [17–18]. The CoLaus study was approved by the Institutional Ethic's Committee of the University of Lausanne. The London Life Sciences Prospective Population Study (LOLIPOP) is an ongoing population-based cohort study of ~30,000 Indian Asian and European white men and women, aged 35–75 years living in West London, United Kingdom [59]. All study participants gave written consent including for genetic studies. The LOLIPOP study is approved by the local Research Ethics Committee. The participants included in the present study are a subset of 3947 Indian Asians and 1601 Europeans from the LOLIPOP cohort study. LOLIPOP individuals are separated by origin as well as the genotyping platform, with IAA, IAI or IAP denoting Indian Asians genotyped on Affymetrix, Illumina or Perlegen platforms, respectively, and EWA and EWI denoting Europeans genotyped on Affymetrix or Illumina platforms, respectively (see Table S1). The InCHIANTI study is a population-based epidemiological study aimed at evaluating the factors that influence mobility in the older population living in the Chianti region in Tuscany, Italy. The details of the study have been previously reported [60]. Overnight fasted blood samples were taken for genomic DNA extraction, and measurement of serum calcium. For this study, 1196 subjects with serum calcium and GWAS data were analyzed. The study protocol was approved by the Italian National Institute of Research and Care of Aging Institutional Review and Medstar Research Institute (Baltimore, MD). The Baltimore longitudinal study on Aging (BLSA) study is a population-based study aimed to evaluate contributors of healthy aging in the older population residing predominantly in the Baltimore-Washington DC area [61]. Starting in 1958, participants are examined every one to four years depending on their age. Blood samples were collected for DNA extraction. This analysis focused on a subset of the participants (N = 717) of European ancestry. The BLSA has continuing approval from the Institutional Review Board (IRB) of Medstar Research Institute. Approval was obtained from local ethic committees for all studies and all participants signed informed written consent. The deCODE study consists of individuals who visited a private outpatient laboratory, the Laboratory in Mjodd, Reykjavik, Iceland between 1997 and 2008. The main referral center for this laboratory is a multispecialty medical clinic in Reykjavik (Laeknasetrid). For the serum calcium analysis we used information on 4,126 individuals with both genome-wide SNP data and measured serum calcium and serum albumin. The samples for bone density analysis have previously been described in detail [24–25]. For this study 6,657 individuals with total hip bone mineral density (BMD) and 6,838 individuals with lumbar spine BMD and SNP data were available for analysis. All participants gave informed consent and the study was approved by the Data Protection Commission of Iceland (DPC) and the National Bioethics Committee of Iceland.

Clinical data

For each CoLaus participant a venous blood sample was collected under fasting conditions. Measurements were conducted using a Modular P apparatus (Roche Diagnostics, Switzerland). Total serum calcium was measured by O-cresolphtalein (2.1% – 1.5% maximum inter and intra-batch CVs); albumin was

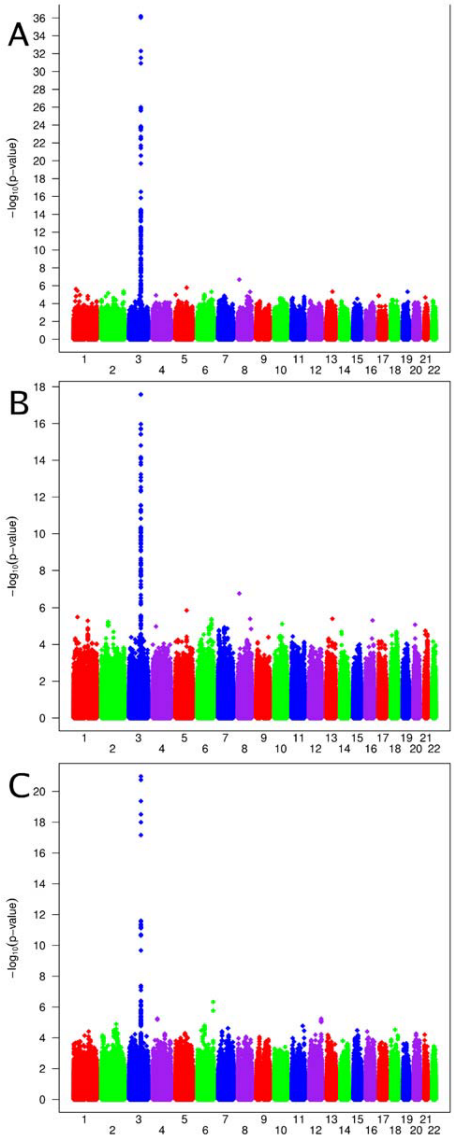


Figure 1. Genome-wide association results. Manhattan plots showing significance of association of all SNPs in the meta-analysis for (A) combined European and Indian Asian cohorts, (B) European cohorts only and (C) Indian Asian cohorts only. SNPs are plotted on the x-axis according to their position on each chromosome against association with serum calcium concentrations on the y-axis (shown as $-\log_{10}$ p-values).
doi:10.1371/journal.pgen.1001035.g001

measured by bromocresol green (2.5% – 0.4%). To further characterize the identified genetic variants, we analyzed the association with several outcomes postulated to be correlated with serum calcium. Within the CoLaus study, we have questionnaire responses to queries about personal histories of osteoporosis, osteoarthritis, myocardial infarction and stroke in addition to clinical data determining hypertension status, defined as previously described [17]. The assessment of LOLIPOP study participants was carried out by a trained research nurse, during a 45 minute appointment according to a standardized protocol and with regular QC audits. An interviewer-administered questionnaire was used to collect data on medical history, family history, current prescribed medication, and cardiovascular risk factors. Physical assessment included anthropometric measurements (height, weight, waist, hip) and blood pressure. Blood was collected after an 8 hour fast for biochemical analysis, including glucose, insulin, total and HDL cholesterol and triglycerides, and whole blood was taken for DNA extraction [59]. InCHIANTI serum albumin concentrations were determined as percentage of total protein using agarose electrophoretic technique (Hydragel Protein (E) 15/30, Sebia, Issy-les-Moulineaux, France). Serum calcium was measured using calorimetric assay (Roch Diagnostic, GmbH, Mannheim, Germany) by a Roche-Hitachi autoanalyzer (The intra-assay CV and 0.9% and the inter-assay CV was 1.5%). Measures of bone density, bone dimensions and osteoporosis diagnosis were assessed by peripheral quantitative computed tomography (pQCT) using the XCT 2000 device (Stratec Medizintechnik, Pforzheim, Germany) [62]. BLSA albumin concentrations were measured by a calorimetric assay using bromocresol green (Ortho-Clinical Diagnostics). Calcium concentrations were measured by a calorimetric assay (Vitros 5,1,FS).

Genome-wide genotyping and imputation

CoLaus participants were genotyped using Affymetrix Human Mapping 500 K Array. For the genome-wide association stage, genotyping in LOLIPOP participants was carried out using the Illumina 317 K mapping array, Affymetrix Human Mapping 500 K array, and Perlegen, 284 K platforms (Table S1). Participants of the InCHIANTI and BLSA studies were genotyped using Illumina Infinium HumanHap 550 K SNP arrays were used for genotyping [21]. Imputation of allele dosage of SNPs was performed using either MACH [63] or IMPUTE [64] with parameters and quality control filters as described in Table S1. All European cohorts imputed SNPs typed in the HapMap CEU population; LOLIPOP Indian Asian cohort imputed SNPs using mixed HapMap populations, given that this showed greater concordance with real genotypes compared with use of any one HapMap population. SNPs were excluded if cohort-specific imputation quality as assessed by r2.hat (MACH) or .info (IMPUTE) metrics were <0.30. In total, 2,557,252 genotyped or imputed SNPs had data from one or more cohorts and were analyzed. Genotypes in deCODE were measured using either humanHap300, humanHap300-duo or humanCNV370.

Statistical analysis

Individual genome-wide association analysis. Biologically active serum calcium is estimated by the correction, Ca_corrected = total serum calcium [mmol/L] + (40 - albumin [g/L])/40. Individuals with values <1.9 or >3 were removed as these were extreme outliers. Linear-regression analyses were carried out using an additive genetic model on log10-transformed corrected calcium levels adjusted for age and pseudosex (a factor variable with three values: males, pre-menopausal females and post-menopausal females). BLSA also included the first two and LOLIPOP included the first four ancestry principal components in the regression, respectively. Regression analyses were performed with QUICKTEST [65] (CoLaus), MACH2qtl (LOLIPOP) [63] or MERLIN [66] (InCHIANTI, BLSA).

Meta-analysis. The results from all cohorts were combined into a fixed-effects meta-analysis using inverse variance weighting. Tests for heterogeneity were assessed using Cochran's Q statistic and the log of the related H statistic [67] after grouping LOLIPOP subsets into European and Indian subsets. For rs1801725 and rs1042636 the p-values were (0.07657, 0.1432) and (0.3450, 0.8876), respectively, indicating

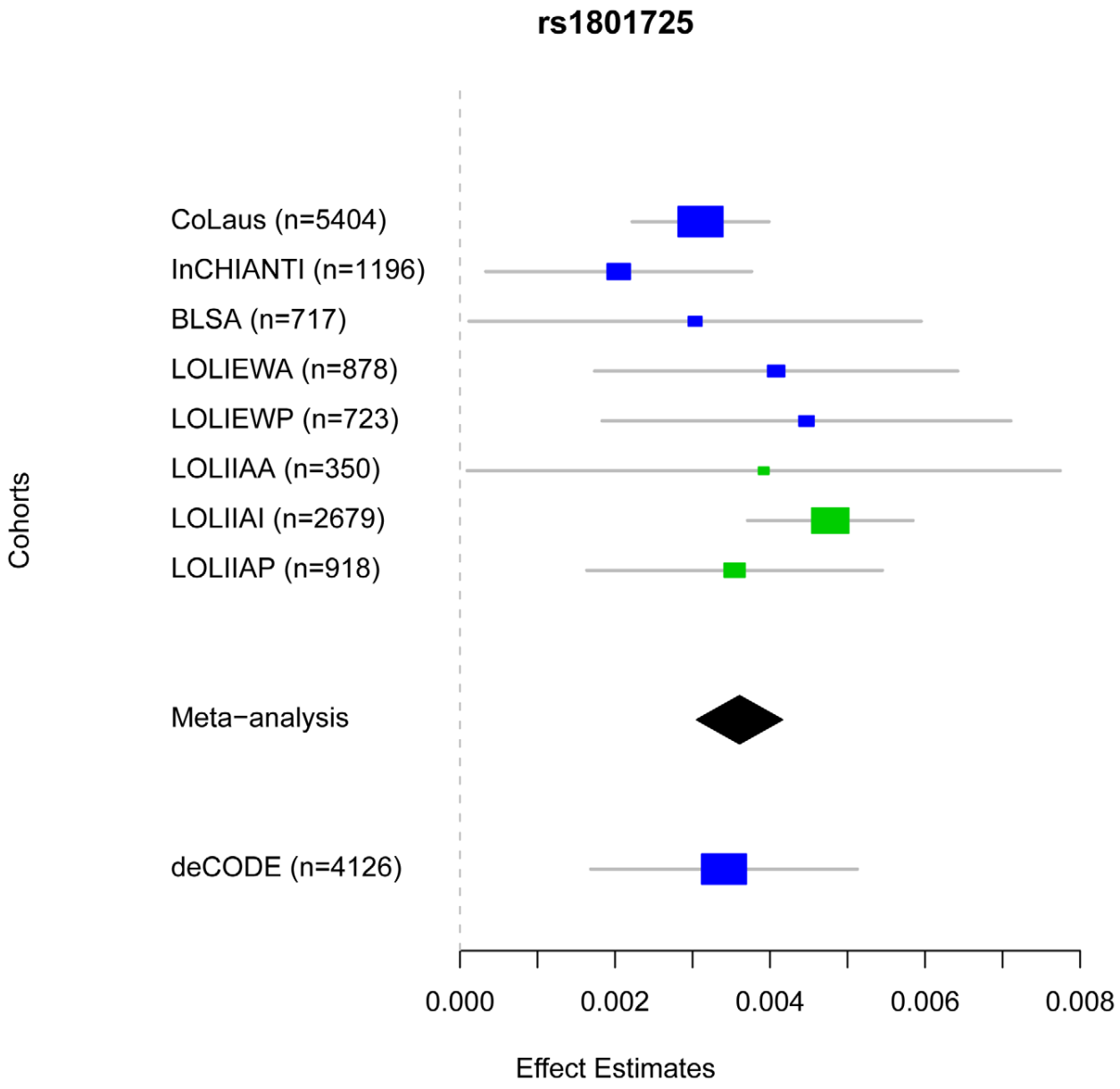


Figure 2. Comparison of rs1801725 significance across cohorts. The effect size and 95% confidence intervals of the serum calcium increasing *T* allele of SNP rs1801725 are shown separately for each cohort (CoLaus, LOLIPOP_EWA, LOLIPOP_EWP, LOLIPOP_IAA, LOLIPOP_IAI, LOLIPOP_IAP, BLSA, InCHIANTI) and for the replication cohort deCODE. European cohorts are shown in blue and Indian Asian cohorts are drawn in green. The size of the box is proportional to the precision $1/se^2$ and the meta-analysis estimate and 95% confidence interval across all cohorts is given by a diamond.
doi:10.1371/journal.pgen.1001035.g002

limited between-study variability. The analysis was implemented in R and run on a quad-core Linux machine. SNPs were reported provided they had effect size estimates in at least 2 of the 5 European cohorts, in at least 2 of the 3 Indian Asian cohorts, or in at least 3 of the 8 total cohorts. For the overall meta-analysis, residual inflation of the test statistic was corrected using genomic control [68]. The inflation factor was 1.0207 for the all combined cohorts, 1.0068 for European cohorts and 1.0286 for Indian Asian cohorts. Where reported, individual study p-values are corrected for inflation using genomic control methods for genotyped and imputed SNPs combined (inflation factors for individual studies were 1.0139 (CoLaus), 0.9891 (LOLIPOP EWA), 0.9994 (LOLIPOP EWP), 0.9967 (LOLIPOP IAA), 1.0131 (LOLIPOP IAI), 0.9985 (LOLIPOP IAP), 0.9842 (InCHIANTI), 1.0019 (BLSA)). The regional association plot (Figure 3) was created modifying a publically available R script [69]. The map of fine-scale recombination rates was downloaded from the HapMap website http://www.hapmap.org/downloads/recombination/ using Phase II HapMap data (release 21). Quantile-quantile plots of the association results are shown in Figure S1A, S1B, S1C, study-specific quantile-quantile plots are shown in Figure S2. Associations below $p = 5 \times 10^{-8}$ were

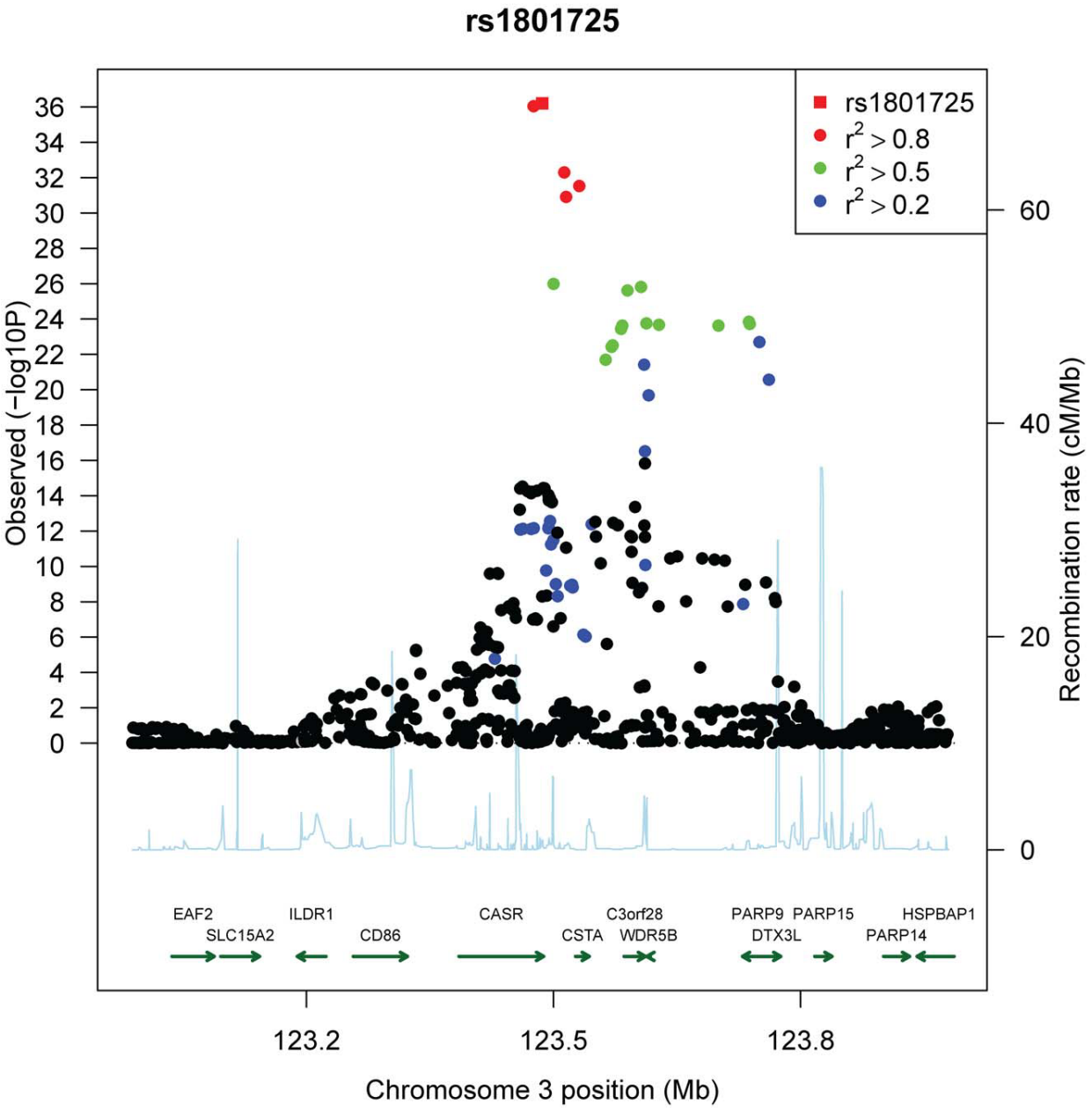


Figure 3. Regional association plot of the *CASR* locus. Plots show genomic position on the x-axis and $-\log_{10}$ p-values on the y-axis for SNPs in the *CASR* locus on chromosome 3. The sentinel hit is shown as a red square. Patterns of linkage disequilibrium between the sentinel SNP and all other SNPs are color-coded. Red circles indicate high correlation ($r^2>0.8$), green circles indicate moderate correlation ($r^2>0.5$), blue circles indicate low correlation ($r^2>0.2$) and black circles indicate no correlation ($r^2<0.2$). The fine-scale recombination rates from [72–73] are plotted in light blue.
doi:10.1371/journal.pgen.1001035.g003

considered genome-wide significant, which corresponds to a Bonferroni correction for the estimated one million independent common variant tests in the human genome of European individuals [70]. The analysis of osteoporosis status in CoLaus and InCHIANTI was performed using logistic regression including age and pseudosex as covariates in QUICKTEST [65]. Linkage disequilibrium was estimated from HapMap CEU (2007-01, build 35 non-redundant) genotypes. LD r^2 statistics were estimated for SNPs within 500 kb using Haploview [71].

Association of rs1801725 with calcium-related outcomes. For each related trait (coronary heart disease, hypertension, kidney stones, myocardial infarction, osteoarthritis, osteoporosis and stroke) we performed a fixed-effects meta-analysis of the logistic regression coefficients. We applied the bonferroni correction to adjust for multiple testing. We performed Wald-based power calculations using a type I error of 0.05/7 and meta-analysis coefficient estimates and standard errors to estimate the sample size for each trait giving power 0.80.

Table 3. Significance of top SNPs.

Combined European and Indian Asian Cohorts

db SNP	Chr	Position (Build 35)	Coded Allele	Non-Coded Allele	Frequency Coded Allele	Beta	Se	Lower 95% CI	Upper 95% CI	P-value	R^2
rs1801725	3	123486447	T	G	17.75	3.61E-03	2.81E-04	3.06E-03	4.16E-03	**6.29E-37**	0.0126
rs17120351	8	14731368	T	C	96.24	4.03E-03	7.67E-04	2.52E-03	5.53E-03	2.06E-07	0.0021
rs7448017	5	117800594	T	G	17.83	-1.73E-03	3.57E-04	-2.43E-03	-1.03E-03	1.65E-06	0.0018
rs742393	1	25271187	G	A	78.86	2.11E-03	4.44E-04	1.24E-03	2.98E-03	2.41E-06	0.0018
rs16827695	1	40956147	G	A	94.52	3.46E-03	7.41E-04	2.01E-03	4.92E-03	3.74E-06	0.0017
rs1550532	2	234046848	G	C	74.24	-1.15E-03	2.48E-04	-1.64E-03	-6.66E-04	4.28E-06	0.0017
rs261503	13	81305121	G	A	91.35	2.45E-03	5.29E-04	1.41E-03	3.49E-03	4.60E-06	0.0017
rs10119	19	50098513	G	A	73.82	-1.63E-03	3.53E-04	-2.33E-03	-9.41E-04	4.76E-06	0.0017
rs17666460	6	149298132	G	A	11.16	1.93E-03	4.18E-04	1.11E-03	2.75E-03	4.81E-06	0.0017
rs16902486	8	129024247	G	C	4.26	-2.66E-03	5.75E-04	-3.78E-03	-1.53E-03	4.83E-06	0.0017
rs17005914	2	70721826	T	C	21.4	-1.47E-03	3.24E-04	-2.11E-03	-8.38E-04	6.76E-06	0.0016
rs10455097	6	74550153	C	A	47.75	-9.58E-04	2.15E-04	-1.38E-03	-5.37E-04	9.94E-06	0.0015

European Cohorts

db SNP	Chr	Position (Build 35)	Coded Allele	Non-Coded Allele	Frequency Coded Allele	Beta	Se	Lower 95% CI	Upper 95% CI	P-value	R^2
rs1801725	3	123486447	T	G	16.76	3.11E-03	3.55E-04	2.41E-03	3.80E-03	**2.58E-18**	0.0085
rs17120351	8	14731368	T	C	96.24	4.03E-03	7.67E-04	2.52E-03	5.53E-03	1.69E-07	0.0031
rs7448017	5	117800594	T	G	17.83	-1.73E-03	3.57E-04	-2.43E-03	-1.03E-03	1.40E-06	0.0026
rs16827695	1	40956147	G	A	94.45	3.46E-03	7.41E-04	2.01E-03	4.92E-03	3.20E-06	0.0024
rs261503	13	81305121	G	A	91.35	2.45E-03	5.29E-04	1.41E-03	3.49E-03	3.95E-06	0.0024
rs16902486	8	129024247	G	C	4.7	-2.92E-03	6.30E-04	-4.15E-03	-1.68E-03	4.04E-06	0.0024
rs17666460	6	149298132	G	A	11.16	1.93E-03	4.18E-04	1.11E-03	2.75E-03	4.13E-06	0.0024
rs12325114	16	72861392	T	C	76.14	-1.40E-03	3.05E-04	-2.00E-03	-8.01E-04	4.86E-06	0.0024
rs6427310	1	150628843	T	C	16.29	-1.65E-03	3.60E-04	-2.35E-03	-9.41E-04	5.09E-06	0.0023
rs17005914	2	70721826	T	C	21.4	-1.47E-03	3.24E-04	-2.11E-03	-8.38E-04	5.83E-06	0.0023
rs12416668	10	83645028	T	C	94.94	-2.77E-03	6.18E-04	-3.98E-03	-1.56E-03	7.60E-06	0.0023
rs6111021	20	16221853	C	A	41.75	1.20E-03	2.67E-04	6.71E-04	1.72E-03	8.36E-06	0.002

Indian Asian Cohorts

db SNP	Chr	Position (Build 35)	Coded Allele	Non-Coded Allele	Frequency Coded Allele	Beta	Se	Lower 95% CI	Upper 95% CI	P-value	R^2
rs17251221	3	123475937	G	A	19.24	4.72E-03	4.86E-04	3.76E-03	5.67E-03	**1.07E-21**	0.0233
rs13203335	6	166194438	T	C	28.04	-2.91E-03	5.70E-04	-4.03E-03	-1.79E-03	4.70E-07	0.0066
rs4695355	4	48030638	G	A	40.26	1.72E-03	3.75E-04	9.89E-04	2.46E-03	5.70E-06	0.0053
rs10846917	12	124568003	T	C	59.69	1.74E-03	3.78E-04	9.98E-04	2.48E-03	5.79E-06	0.0053

We report SNPs with p-values <1E-05 filtered by distinct regions, determined by merging SNPs within 1 Mb of each other. Results are given for different cohort subsets of combined European and Indian Asian cohorts, European cohorts, and Indian Asian cohorts. The Beta and SE values represent the effect of the coded allele on $\log_{10}$ corrected serum calcium levels; R^2 represents the fraction of variation explained by the SNP.

doi:10.1371/journal.pgen.1001035.t003

Supporting Information

Figure S1 Quantile-Quantile plots of genome-wide association results. Observed -$\log_{10}$ p-values on the y-axis are plotted against theoretical -$\log_{10}$ p-values on the x-axis resulting from each meta-analysis. Results are color-coded by chromosome. The top results largely derive from the CASR locus on chromosome 3. Results are shown separately for (A) all cohorts, (B) European cohorts and (C) Indian Asian cohorts.

Found at: doi:10.1371/journal.pgen.1001035.s001 (6.48 MB TIF)

Figure S2 Study-specific quantile-quantile plots. Shown are observed -$\log_{10}$ p-values plotted against expected -$\log_{10}$ p-values resulting from each single study after applying genomic control correction. The study-specific λ-values were λ = 1.0139 (CoLaus), λ = 0.9891 (LOLIPOP_EWA), λ = 0.9994 (LOLIPOP_EWP), λ = 0.9967 (LOLIPOP_IAA), λ = 1.0131 (LOLIPOP_IAI), λ = 0.9985 (LOLIPOP_IAP), λ = 0.9842 (InCHIANTI),

$\lambda = 1.0019$ (BLSA). For the combined European and Indian Asian, European only and Indian Asian only meta-analyses the inflation factors were 1.0207, 1.0068, and 1.0286, respectively.
Found at: doi:10.1371/journal.pgen.1001035.s002 (4.36 MB TIF)

Figure S3 Comparison of significance across cohorts. The effect size and 95% confidence intervals of SNPs which do not reach genome-wide significance in the combined European and Indian Asian meta-analysis are shown separately for each cohort (CoLaus, LOLIPOP_EWA, LOLIPOP_EWP, LOLIPOP_IAA, LOLIPOP_IAI, LOLIPOP_IAP, BLSA, InCHIANTI). European cohorts are drawn in blue and Indian Asian cohorts are drawn in green. The size of the box is proportional to the precision $1/se^2$ and the meta-analysis estimate and 95% confidence interval across all cohorts is given by a diamond.
Found at: doi:10.1371/journal.pgen.1001035.s003 (0.38 MB TIF)

Figure S4 Candidate gene QQ-plots. For 18611 SNPs mapping to candidate genes (excluding CASR), we compare observed $-\log_{10}$ p-values to the mean quantiles of the uniform distribution. As a comparison, we randomly choose a set of genes from which we select the same number of SNPs. From 1,000 random draws we calculate the 95th percentile of $-\log_{10}$ p-values (in blue). Results are shown separately for all cohorts, European only and Indian Asian only (A–C). CoLaus permuted phenotype results comparing observed p-values to the 95th percentile of $-\log_{10}$ p-values from 100 permutations are shown in (D).
Found at: doi:10.1371/journal.pgen.1001035.s004 (0.41 MB TIF)

Table S1 Genotyping, imputation, and analysis procedures by study. The genotyping platforms, quality control (QC) filters applied before imputation, imputation software, number of SNPs, and genotype-phenotype association software are shown for each study.
Found at: doi:10.1371/journal.pgen.1001035.s005 (0.05 MB DOC)

Table S2 Complete list of genome-wide significant SNPs. Below is the list of all SNPs that exceeded the threshold of genome-wide significance ($p<5\times10e-8$). Position is given for NCBI Build 35. Meta-analysis is performed by inverse variance weighted fixed effect regression. The coded allele is the allele to which the beta (effect) estimate refers. Results are shown separately for (A) European and Indian Asian cohorts, (B) European cohorts only, and (C) Indian Asian cohorts only.
Found at: doi:10.1371/journal.pgen.1001035.s006 (0.24 MB DOC)

Table S3 Significance of top SNPs by cohort. Shown are study-specific results of the SNPs with genomic control (GC) p-values <1E-05 filtered by distinct regions, determined by merging SNPs within 1 Mb of each other. Results are shown separately for (A) European and Indian Asian cohorts, (B) European cohorts, and (C) Indian Asian cohorts.
Found at: doi:10.1371/journal.pgen.1001035.s007 (0.08 MB DOC)

Table S4 Number of cases and controls for calcium-related outcomes. For several related phenotypes, we test the association of rs1801725 with these binary responses. Shown here are the number of cases and controls for each phenotype in each cohort and the total across cohorts.
Found at: doi:10.1371/journal.pgen.1001035.s008 (0.04 MB DOC)

Table S5 Logistic regression of clinical phenotypes on rs1801725. We report the effect size and standard error of the rs1801725 T allele from logistic regressions of each clinical phenotype.
Found at: doi:10.1371/journal.pgen.1001035.s009 (0.04 MB DOC)

Table S6 Studies of CASR mutations and serum calcium. A survey of previous studies which investigate the relationship between CASR mutations and levels of serum calcium.
Found at: doi:10.1371/journal.pgen.1001035.s010 (0.05 MB DOC)

Acknowledgments

The authors would like to thank those who agreed to participate in the studies and the many colleagues who contributed to collection and phenotypic characterization of the clinical samples, to genotyping, and to statistical analyses. The computations for CoLaus imputation were performed in part at the Vital-IT center for high performance computing of the Swiss Institute of Bioinformatics.

Author Contributions

Conceived and designed the experiments: K Kapur, R DeSilva, P Elliott, JM Guralnik, J Scott, S Bandinelli, J Chambers, G Waeber, L Ferrucci, JS Kooner, V Mooser, P Vollenweider, JS Beckmann, M Bochud, S Bergmann. Analyzed the data: T Johnson, ND Beckmann, J Sehmi, T Tanaka, Z Kutalik, U Styrkarsdottir, W Zhang, DF Gudbjartsson. Contributed reagents/materials/analysis tools: D Marek, Y Milaneschi, H Holm, A DiIorio, D Waterworth, Y Li, AB Singleton, US Bjornsdottir, G Sigurdsson, DG Hernandez, R DeSilva, P Elliott, GI Eyjolfsson, JM Guralnik, J Scott, U Thorsteinsdottir, S Bandinelli, K Stefansson. Wrote the paper: K Kapur, JS Beckmann, M Bochud, S Bergmann.

References

1. Carafoli E (2004) Calcium-mediated cellular signals: a story of failures. Trends Biochem Sci 29: 371–379.
2. Carafoli E (2005) Calcium–a universal carrier of biological signals. Delivered on 3 July 2003 at the Special FEBS Meeting in Brussels. FEBS J 272: 1073–1089.
3. Carafoli E (2004) The ambivalent nature of the calcium signal. J Endocrinol Invest 27: 134–136.
4. Whitfield JB, Martin NG (1984) The effects of inheritance on constituents of plasma: a twin study on some biochemical variables. Ann Clin Biochem 21 (Pt 3): 176–183.
5. Williams PD, Puddey IB, Martin NG, Beilin LJ (1992) Platelet cytosolic free calcium concentration, total plasma calcium concentration and blood pressure in human twins: a genetic analysis. Clin Sci (Lond) 82: 493–504.
6. Leifsson BG, Ahren B (1996) Serum calcium and survival in a large health screening program. J Clin Endocrinol Metab 81: 2149–2153.
7. Lind L, Skarfors E, Berglund L, Lithell H, Ljunghall S (1997) Serum calcium: a new, independent, prospective risk factor for myocardial infarction in middle-aged men followed for 18 years. J Clin Epidemiol 50: 967–973.
8. Dhingra R, Sullivan LM, Fox CS, Wang TJ, D'Agostino RB, Sr, et al. (2007) Relations of serum phosphorus and calcium levels to the incidence of cardiovascular disease in the community. Arch Intern Med 167: 879–885.
9. Palmer M, Adami HO, Bergstrom R, Jakobsson S, Akerstrom G, et al. (1987) Survival and renal function in untreated hypercalcaemia. Population-based cohort study with 14 years of follow-up. Lancet 1: 59–62.
10. Lundgren E, Lind L, Palmer M, Jakobsson S, Ljunghall S, et al. (2001) Increased cardiovascular mortality and normalized serum calcium in patients with mild hypercalcemia followed up for 25 years. Surgery 130: 978–985.
11. Wermers RA, Khosla S, Atkinson EJ, Grant CS, Hodgson SF, et al. (1998) Survival after the diagnosis of hyperparathyroidism: a population-based study. Am J Med 104: 115–122.
12. Heath H, III, Odelberg S, Jackson CE, Teh BT, Hayward N, et al. (1996) Clustered inactivating mutations and benign polymorphisms of the calcium receptor gene in familial benign hypocalciuric hypercalcemia suggest receptor functional domains. J Clin Endocrinol Metab 81: 1312–1317.
13. Pollak MR, Brown EM, Chou YH, Hebert SC, Marx SJ, et al. (1993) Mutations in the human Ca(2+)-sensing receptor gene cause familial hypocalciuric hypercalcemia and neonatal severe hyperparathyroidism. Cell 75: 1297–1303.
14. Pollak MR, Brown EM, Estep HL, McLaine PN, Kifor O, et al. (1994) Autosomal dominant hypocalcaemia caused by a Ca(2+)-sensing receptor gene mutation. Nat Genet 8: 303–307.

15. Cole DE, Peltekova VD, Rubin LA, Hawker GA, Vieth R, et al. (1999) A986S polymorphism of the calcium-sensing receptor and circulating calcium concentrations. Lancet 353: 112–115.
16. Cole DE, Vieth R, Trang HM, Wong BY, Hendy GN, et al. (2001) Association between total serum calcium and the A986S polymorphism of the calcium-sensing receptor gene. Mol Genet Metab 72: 168–174.
17. Firmann M, Mayor V, Marques-Vidal PM, Bochud M, Pecoud A, et al. (2008) The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc Disord.
18. Rodondi N, Cornuz J, Marques-Vidal P, Butler J, Hayoz D, et al. (2008) Aspirin use for the primary prevention of coronary heart disease: A population-based study in Switzerland. Prev Med 46: 137–144.
19. Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, et al. (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 40: 716–718.
20. Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, et al. (2008) Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat Genet 40: 149–151.
21. Melzer D, Perry JR, Hernandez D, Corsi AM, Stevens K, et al. (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet 4: e1000072. doi:10.1371/journal.pgen.1000072.
22. Schock NW, Greulich RC, Andres R, Arenberg D, Costa PT, et al. (1984) Normal Human Aging: The Baltimore Longitudinal Study of Aging. Washington D.C.
23. Scillitani A, Guarnieri V, De GS, Muscarella LA, Battista C, et al. (2004) Blood ionized calcium is associated with clustered polymorphisms in the carboxyl-terminal tail of the calcium-sensing receptor. J Clin Endocrinol Metab 89: 5634–5638.
24. Styrkarsdottir U, Halldorsson BV, Gretarsdottir S, Gudbjartsson DF, Walters GB, et al. (2009) New sequence variants associated with bone mineral density. Nat Genet 41: 15–17.
25. Styrkarsdottir U, Halldorsson BV, Gretarsdottir S, Gudbjartsson DF, Walters GB, et al. (2008) Multiple genetic loci for bone mineral density and fractures. N Engl J Med 358: 2355–2365.
26. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, et al. (2008) Many sequence variants affecting diversity of adult human height. Nat Genet 40: 609–615.
27. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, et al. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40: 584–591.
28. Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, et al. (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Genet 39: 1245–1250.
29. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40: 575–583.
30. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889–894.
31. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40: 768–775.
32. Doring A, Gieger C, Mehta D, Gohlke H, Prokisch H, et al. (2008) SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. Nat Genet 40: 430–436.
33. Li S, Sanna S, Maschio A, Busonero F, Usala G, et al. (2007) The GLUT9 gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. PLoS Genet 3: e194. doi:10.1371/journal.pgen.0030194.
34. Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, et al. (2008) Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. Am J Hum Genet 82: 139–149.
35. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, et al. (2008) LDL-cholesterol concentrations: a genome-wide association study. Lancet 371: 483–491.
36. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161–169.
37. Goldstein DB (2009) Common genetic variation and human traits. N Engl J Med 360: 1696–1698.
38. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.
39. Bai M, Trivedi S, Brown EM (1998) Dimerization of the extracellular calcium-sensing receptor (CaR) on the cell surface of CaR-transfected HEK293 cells. J Biol Chem 273: 23605–23610.
40. Bai M, Trivedi S, Lane CR, Yang Y, Quinn SJ, et al. (1998) Protein kinase C phosphorylation of threonine at position 888 in Ca2+o-sensing receptor (CaR) inhibits coupling to Ca2+ store release. J Biol Chem 273: 21267–21275.
41. Gama L, Breitwieser GE (1998) A carboxyl-terminal domain controls the cooperativity for extracellular Ca2+ activation of the human calcium sensing receptor. A study with receptor-green fluorescent protein fusions. J Biol Chem 273: 29712–29718.
42. Vezzoli G, Terranegra A, Arcidiacono T, Biasion R, Coviello D, et al. (2007) R990G polymorphism of calcium-sensing receptor does produce a gain-of-function and predispose to primary hypercalciuria. Kidney Int 71: 1155–1162.
43. Harding B, Curley AJ, Hannan FM, Christie PT, Bowl MR, et al. (2006) Functional characterization of calcium sensing receptor polymorphisms and absence of association with indices of calcium homeostasis and bone mineral density. Clin Endocrinol (Oxf) 65: 598–605.
44. Kelly C, Gunn IR, Gaffney D, Devgun MS (2006) Serum calcium, urine calcium and polymorphisms of the calcium sensing receptor gene. Ann Clin Biochem 43: 503–506.
45. Laaksonen MM, Outila TA, Karkkainen MU, Kemi VE, Rita HJ, et al. (2009) Associations of vitamin D receptor, calcium-sensing receptor and parathyroid hormone gene polymorphisms with calcium homeostasis and peripheral bone density in adult Finns. J NutrigenetNutrigenomics 2: 55–63.
46. Bollerslev J, Wilson SG, Dick IM, Devine A, Dhaliwal SS, et al. (2004) Calcium-sensing receptor gene polymorphism A986S does not predict serum calcium level, bone mineral density, calcaneal ultrasound indices, or fracture rate in a large cohort of elderly women. Calcif Tissue Int 74: 12–17.
47. Cetani F, Borsari S, Vignali E, Pardi E, Picone A, et al. (2002) Calcium-sensing receptor gene polymorphisms in primary hyperparathyroidism. J Endocrinol Invest 25: 614–619.
48. Yamauchi M, Sugimoto T, Yamaguchi T, Yano S, Kanzawa M, et al. (2001) Association of polymorphic alleles of the calcium-sensing receptor gene with the clinical severity of primary hyperparathyroidism. Clin Endocrinol (Oxf) 55: 373–379.
49. Rothe HM, Shapiro WB, Sun WY, Chou SY (2005) Calcium-sensing receptor gene polymorphism Arg990Gly and its possible effect on response to cinacalcet HCl. Pharmacogenet Genomics 15: 29–34.
50. Scillitani A, Guarnieri V, Battista C, De GS, Muscarella LA, et al. (2007) Primary hyperparathyroidism and the presence of kidney stones are associated with different haplotypes of the calcium-sensing receptor. J Clin Endocrinol Metab 92: 277–283.
51. Miedlich S, Lamesch P, Mueller A, Paschke R (2001) Frequency of the calcium-sensing receptor variant A986S in patients with primary hyperparathyroidism. Eur J Endocrinol 145: 421–427.
52. Vezzoli G, Tanini A, Ferrucci L, Soldati L, Bianchin C, et al. (2002) Influence of calcium-sensing receptor gene on urinary calcium excretion in stone-forming patients. J Am Soc Nephrol 13: 2517–2523.
53. Kim KS, Kim GS, Hwang JY, Lee HJ, Park MH, et al. (2007) Single nucleotide polymorphisms in bone turnover-related genes in Koreans: ethnic differences in linkage disequilibrium and haplotype. BMC Med Genet 8: 70.
54. Marz W, Seelhorst U, Wellnitz B, Tiran B, Obermayer-Pietsch B, et al. (2007) Alanine to serine polymorphism at position 986 of the calcium-sensing receptor associated with coronary heart disease, myocardial infarction, all-cause, and cardiovascular mortality. J Clin Endocrinol Metab 92: 2363–2369.
55. Richards JB, Kavvoura FK, Rivadeneira F, Styrkarsdottir U, Estrada K, et al. (2009) Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. Ann Intern Med 151: 528–537.
56. Björkman MP, Sorva AJ, Tilvis RS (1979) Calculated serum calcium is an insufficient surrogate for measured ionized calcium. Archives of Gerontology and Geriatrics 49: 348–350.
57. Robertson WG, Marshall RW (1979) Calcium measurements in serum and plasma–total and ionized. CRC Crit Rev Clin Lab Sci 11: 271–304.
58. Ladenson JH, Lewis JW, Boyd JC (1978) Failure of total calcium corrected for protein, albumin, and pH to correctly assess free calcium status. J Clin Endocrinol Metab 46: 986–993.
59. Chambers JC, Zhang W, Li Y, Sehmi J, Wass MN, et al. (2009) Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nature Genetics.
60. Ferrucci L, Bandinelli S, Benvenuti E, Di Iorio A, Macchi C, et al. (2000) Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J Am Geriatr Soc 48: 1618–1625.
61. Shock NW, Greulich RC, Arenberg D, Costa PT, Lakatta EG, et al. (1984) Normal Human Aging: The Baltimore Longitudinal Study of Aging. Washington, D.C.: National Institutes of Health.
62. Russo CR, Lauretani F, Bandinelli S, Bartali B, Di Iorio A, et al. (2003) Aging bone in men and women: beyond changes in bone mineral density. Osteoporos Int 14: 531–538.
63. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10: 387–406.
64. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906–913.
65. Johnson T, Kutalik Z (2008) QUICKTEST.
66. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101.
67. Higgins JP, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Stat Med 21: 1539–1558.
68. Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66: 1933–1944.

69. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331–1336.
70. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32: 227–234.
71. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
72. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304: 581–584.
73. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, et al. (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107–111.

nature genetics

Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index

Obesity is globally prevalent and highly heritable, but its underlying genetic factors remain largely elusive. To identify genetic loci for obesity susceptibility, we examined associations between body mass index and ~2.8 million SNPs in up to 123,865 individuals with targeted follow up of 42 SNPs in up to 125,931 additional individuals. We confirmed 14 known obesity susceptibility loci and identified 18 new loci associated with body mass index ($P < 5 \times 10^{-8}$), one of which includes a copy number variant near *GPRC5B*. Some loci (at *MC4R*, *POMC*, *SH2B1* and *BDNF*) map near key hypothalamic regulators of energy balance, and one of these loci is near *GIPR*, an incretin receptor. Furthermore, genes in other newly associated loci may provide new insights into human body weight regulation.

Obesity is a major and increasingly prevalent risk factor for multiple disorders, including type 2 diabetes and cardiovascular disease[1,2]. Although lifestyle changes have driven its prevalence to epidemic proportions, heritability studies provide evidence for a substantial genetic contribution (with heritability estimates (h^2) of ~40%–70%) to obesity risk[3,4]. BMI is an inexpensive, non-invasive measure of obesity that predicts the risk of related complications[5]. Identifying genetic determinants of BMI could lead to a better understanding of the biological basis of obesity.

Genome-wide association studies (GWAS) of BMI have previously identified ten loci with genome-wide significant ($P < 5 \times 10^{-8}$) associations[6–10] in or near *FTO, MC4R*, *TMEM18*, *GNPDA2*, *BDNF*, *NEGR1*, *SH2B1*, *ETV5*, *MTCH2* and *KCTD15*. Many of these genes are expressed or known to act in the central nervous system, highlighting a likely neuronal component in the predisposition to obesity[9]. This pattern is consistent with results in animal models and studies of monogenic human obesity in which neuronal genes, particularly those expressed in the hypothalamus and involved in regulation of appetite or energy balance, are known to play a major role in susceptibility to obesity[11–13].

The ten previously identified loci account for only a small fraction of the variation in BMI. Furthermore, power calculations based on the effect sizes of established variants have suggested that increasing the sample size would likely lead to the discovery of additional variants[9]. To identify additional loci associated with BMI, we expanded the Genetic Investigation of Anthropometric Traits (GIANT) Consortium genome-wide association meta-analysis to include a total of 249,796 individuals of European ancestry.

RESULTS

Stage 1 GWAS identifies new loci associated with BMI

We first conducted a meta-analysis of GWAS of BMI and ~2.8 million imputed or genotyped SNPs using data from 46 studies including up to 123,865 individuals (Online Methods, **Supplementary Fig. 1** and **Supplementary Note**). This stage 1 analysis revealed 19 loci associated with BMI at $P < 5 \times 10^{-8}$ (**Table 1**, **Fig. 1a** and **Supplementary Table 1**). These 19 loci included all ten loci from previous GWAS of BMI[6–10], two loci previously associated with body weight[10] (at *FAIM2* and *SEC16B*) and one locus previously associated with waist circumference[14] (near *TFAP2B*). The remaining six loci, near *GPRC5B*, *MAP2K5-LBXCOR1*, *TNNI3K*, *LRRN6C*, *FLJ35779-HMGCR* and *PRKD1*, have not previously been associated with BMI or other obesity-related traits.

Stage 2 follow up identifies additional new loci for BMI

To identify additional BMI-associated loci and to validate the loci that reached genome-wide significance in the stage 1 analyses, we examined SNPs representing 42 independent loci (including the 19 genome-wide significant loci) having a stage 1 $P < 5 \times 10^{-6}$. Variants were considered to be independent if the pair-wise linkage disequilibrium (LD, r^2) was less than 0.1 and if they were separated by at least 1 Mb. In stage 2, we examined these 42 SNPs in up to 125,931 additional individuals (79,561 newly genotyped individuals from 16 different studies and 46,370 individuals from 18 additional studies for which genome-wide association data were available; **Table 1**, **Supplementary Note** and Online Methods). In a joint analysis of stage 1 and stage 2 results, 32 of the 42 SNPs reached $P < 5 \times 10^{-8}$ (**Table 1**, **Supplementary Table 1** and **Supplementary Figs. 1** and **2**). Even after excluding SNPs within the 32 confirmed BMI loci, we still observed an excess of small *P* values compared to the distribution expected under the null hypothesis (**Fig. 1b** and **Supplementary Fig. 3**), suggesting that more BMI loci remain to be uncovered.

The 32 confirmed associations included all 19 loci with $P < 5 \times 10^{-8}$ at stage 1, 12 additional new loci near *RBJ-ADCY3-POMC, QPCTL-GIPR*, *SLC39A8*, *TMEM160*, *FANCL*, *CADM2*, *LRP1B*, *PTBP2*, *MTIF3-GTF3A*, *ZNF608*, *RPL27A-TUB* and *NUDT3-HMGA1* and one locus (in *NRXN3*) previously associated with waist circumference[15] (**Table 1**, **Supplementary Table 1** and **Supplementary Figs. 1** and **2**). In all, our study increased the number of loci robustly associated with BMI from 10 to 32. Four of the 22 new loci were previously associated

A full list of authors and affiliations appear at the end of the paper.

Received 13 May; accepted 15 September; published online 10 October 2010; doi:10.1038/ng.686

with body weight[10] or waist circumference[14,15], whereas 18 new loci had not previously associated with any obesity-related trait in the general population. Although we confirmed all loci previously established by large-scale GWAS for BMI[6–10] and waist circumference[14,15], four loci previously identified in GWAS for early-onset or adult morbid obesity[16,17] (at *NPC1*, rs1805081, $P = 0.0025$; *MAF*, rs1424233, $P = 0.25$; *PTER*, rs10508503, $P = 0.64$; and *TNKS-MSRA*, rs473034, $P = 0.23$) showed limited or no evidence of association with BMI in our study.

Table 1 Stage 1 and stage 2 results of the 32 SNPs that were associated with BMI at genome-wide significant ($P < 5 \times 10^{-8}$) levels

SNP	Nearest gene	Other nearby genes[a]	Chr.	Position[b] (bp)	Alleles[b] Effect	Other	Frequency effect allele	Per allele change in BMI β (s.e.m.)[c]	Explained variance (%)	Stage 1 *P*	Stage 2 *P*	Stage 1 + 2 *n*	Stage 1 + 2 *P*
Previously identified BMI loci													
rs1558902	*FTO*		16	52,361,075	A	T	0.42	0.39 (0.02)	0.34%	2.05×10^{-62}	1.01×10^{-60}	192,344	4.8×10^{-120}
rs2867125	*TMEM18*		2	612,827	C	T	0.83	0.31 (0.03)	0.15%	2.42×10^{-22}	4.42×10^{-30}	197,806	2.77×10^{-49}
rs571312	*MC4R* (B)		18	55,990,749	A	C	0.24	0.23 (0.03)	0.10%	1.82×10^{-22}	3.19×10^{-21}	203,600	6.43×10^{-42}
rs10938397	*GNPDA2*		4	44,877,284	G	A	0.43	0.18 (0.02)	0.08%	4.35×10^{-17}	1.45×10^{-15}	197,008	3.78×10^{-31}
rs10767664	*BDNF* (B,M)		11	27,682,562	A	T	0.78	0.19 (0.03)	0.07%	5.53×10^{-13}	1.17×10^{-14}	204,158	4.69×10^{-26}
rs2815752	*NEGR1* (C,Q)		1	72,585,028	A	G	0.61	0.13 (0.02)	0.04%	1.17×10^{-14}	2.29×10^{-9}	198,380	1.61×10^{-22}
rs7359397	*SH2B1* (Q,B,M)	*APOB48R* (Q,M), *SULT1A2* (Q,M), *AC138894.2* (M), *ATXN2L* (M), *TUFM* (Q)	16	28,793,160	T	C	0.40	0.15 (0.02)	0.05%	1.75×10^{-10}	7.89×10^{-12}	204,309	1.88×10^{-20}
rs9816226	*ETV5*		3	187,317,193	T	A	0.82	0.14 (0.03)	0.03%	7.61×10^{-14}	1.15×10^{-6}	196,221	1.69×10^{-18}
rs3817334	*MTCH2* (Q,M)	*NDUFS3* (Q), *CUGBP1* (Q)	11	47,607,569	T	C	0.41	0.06 (0.02)	0.01%	4.79×10^{-11}	1.10×10^{-3}	191,943	1.59×10^{-12}
rs29941	*KCTD15*		19	39,001,372	G	A	0.67	0.06 (0.02)	0.00%	1.31×10^{-9}	2.40×10^{-2}	192,872	3.01×10^{-9}
Previously identified waist and weight loci													
rs543874	*SEC16B*		1	176,156,103	G	A	0.19	0.22 (0.03)	0.07%	1.66×10^{-13}	2.41×10^{-11}	179,414	3.56×10^{-23}
rs987237	*TFAP2B*		6	50,911,009	G	A	0.18	0.13 (0.03)	0.03%	5.97×10^{-16}	2.40×10^{-6}	195,776	2.90×10^{-20}
rs7138803	*FAIM2*		12	48,533,735	A	G	0.38	0.12 (0.02)	0.04%	3.96×10^{-11}	7.82×10^{-8}	200,064	1.82×10^{-17}
rs10150332	*NRXN3*		14	79,006,717	C	T	0.21	0.13 (0.03)	0.02%	2.03×10^{-7}	2.86×10^{-5}	183,022	2.75×10^{-11}
Newly identified BMI loci													
rs713586	*RBJ*	*ADCY3* (Q, M), *POMC* (Q,B)	2	25,011,512	C	T	0.47	0.14 (0.02)	0.06%	1.80×10^{-7}	1.44×10^{-16}	230,748	6.17×10^{-22}
rs12444979	*GPRC5B* (C,Q)	*IQCK* (Q)	16	19,841,101	C	T	0.87	0.17 (0.03)	0.04%	4.20×10^{-11}	8.13×10^{-12}	239,715	2.91×10^{-21}
rs2241423	*MAP2K5*	*LBXCOR1* (M)	15	65,873,892	G	A	0.78	0.13 (0.02)	0.03%	1.15×10^{-10}	1.59×10^{-9}	227,950	1.19×10^{-18}
rs2287019	*QPCTL*	*GIPR* (B,M)	19	50,894,012	C	T	0.80	0.15 (0.03)	0.04%	3.18×10^{-7}	1.40×10^{-10}	194,564	1.88×10^{-16}
rs1514175	*TNNI3K*		1	74,764,232	A	G	0.43	0.07 (0.02)	0.02%	1.36×10^{-9}	7.04×10^{-6}	227,900	8.16×10^{-14}
rs13107325	*SLC39A8* (Q,M)		4	103,407,732	T	C	0.07	0.19 (0.04)	0.03%	1.37×10^{-7}	1.93×10^{-7}	245,378	1.50×10^{-13}
rs2112347	*FLJ35779* (M)	*HMGCR* (B)	5	75,050,998	T	G	0.63	0.10 (0.02)	0.02%	4.76×10^{-8}	8.29×10^{-7}	231,729	2.17×10^{-13}
rs10968576	*LRRN6C*		9	28,404,339	G	A	0.31	0.11 (0.02)	0.02%	1.88×10^{-8}	3.19×10^{-6}	216,916	2.65×10^{-13}
rs3810291	*TMEM160* (Q)	*ZC3H4* (Q)	19	52,260,843	A	G	0.67	0.09 (0.02)	0.02%	1.04×10^{-7}	1.59×10^{-6}	233,512	1.64×10^{-12}
rs887912	*FANCL*		2	59,156,381	T	C	0.29	0.10 (0.02)	0.03%	2.69×10^{-6}	1.72×10^{-7}	242,807	1.79×10^{-12}
rs13078807	*CADM2*		3	85,966,840	G	A	0.20	0.10 (0.02)	0.02%	9.81×10^{-8}	5.32×10^{-5}	237,404	3.94×10^{-11}
rs11847697	*PRKD1*		14	29,584,863	T	C	0.04	0.17 (0.05)	0.01%	1.11×10^{-8}	2.25×10^{-4}	241,667	5.76×10^{-11}
rs2890652	*LRP1B*		2	142,676,401	C	T	0.18	0.09 (0.03)	0.02%	2.38×10^{-7}	9.47×10^{-5}	209,068	1.35×10^{-10}
rs1555543	*PTBP2*		1	96,717,385	C	A	0.59	0.06 (0.02)	0.01%	7.65×10^{-7}	4.48×10^{-5}	243,013	3.68×10^{-10}
rs4771122	*MTIF3*	*GTF3A* (Q)	13	26,918,180	G	A	0.24	0.09 (0.03)	0.02%	1.20×10^{-7}	8.24×10^{-4}	198,577	9.48×10^{-10}
rs4836133	*ZNF608*		5	124,360,002	A	C	0.48	0.07 (0.02)	0.01%	7.04×10^{-7}	1.88×10^{-4}	241,999	1.97×10^{-9}
rs4929949	*RPL27A*	*TUB* (B)	11	8,561,169	C	T	0.52	0.06 (0.02)	0.01%	7.57×10^{-8}	1.00×10^{-3}	249,791	2.80×10^{-9}
rs206936	*NUDT3*	*HMGA1* (B)	6	34,410,847	G	A	0.21	0.06 (0.02)	0.01%	2.81×10^{-6}	7.39×10^{-4}	249,777	3.02×10^{-8}

Chr., chromosome; Q, association and eQTL data converge to affect gene expression; B, biological candidate; M, BMI-associated variant is in strong LD ($r^2 \geq 0.75$) with a missense variant in the indicated gene; C, CNV.
[a]Genes within ± 500 kb of the lead SNP. [b]Positions according to Build 36 and allele coding based on the positive strand. [c]Effect sizes in kg/m^2 obtained from stage 2 cohorts only.

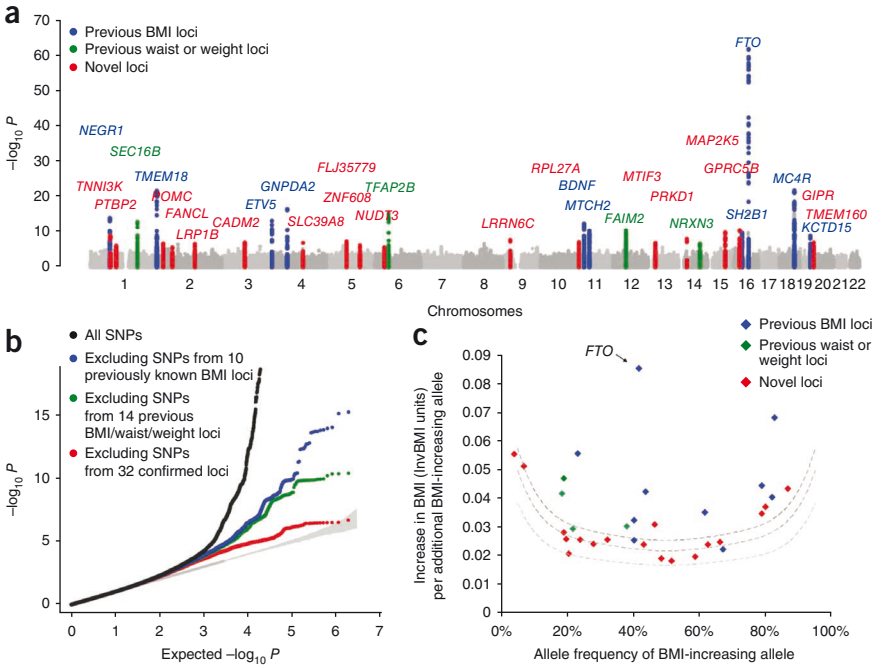


Figure 1 Genome-wide association results for the BMI meta-analysis. (**a**) Manhattan plot showing the significance of association between all SNPs and BMI in the stage 1 meta-analysis, highlighting SNPs previously reported to show genome-wide significant association with BMI (blue), weight or waist circumference (green) and the 18 new regions described here (red). The 19 SNPs that reached genome-wide significance in stage 1 (13 previously reported and 6 new SNPs) are listed in **Table 1**. (**b**) Quantile-quantile plot of SNPs in the stage 1 meta-analysis (black) and after removing any SNPs within 1 Mb of the ten previously reported genome-wide significant hits for BMI (blue), after additionally excluding SNPs from the four loci for waist or weight (green), and after excluding SNPs from all 32 confirmed loci (red). The plot is abridged at the *y* axis (at $P < 10^{-20}$) to better visualize the excess of small *P* values after excluding the 32 confirmed loci (**Supplementary Fig. 3** shows the full-scale quantile-quantile plot). The shaded region is the 95% concentration band. (**c**) Plot of effect size (in inverse-normally transformed units (invBMI)) versus effect-allele frequency of newly identified and previously identified BMI variants after stage 1 and stage 2 meta-analysis, including the 10 previously identified BMI loci (blue), the 4 previously identified waist and weight loci (green) and the 18 newly identified BMI loci (blue). The dotted lines represent the minimum effect sizes that could be identified for a given effect-allele frequency with 80% (upper line), 50% (middle line) and 10% (lower line) power, assuming a sample size of 123,000 individuals and an α level of 5×10^{-8}.

As could be expected, the effect sizes of the 18 newly discovered loci are slightly smaller, for a given minor allele frequency, than those of the previously identified variants (**Table 1** and **Fig. 1c**). The increased sample size used here also brought out more signals with low minor allele frequency. The BMI-increasing allele frequencies for the 18 newly identified variants ranged from 4% to 87%, covering more of the allele frequency spectrum than previous, smaller GWAS of BMI (24%–83%)[9,10] (**Table 1** and **Fig. 1c**).

We tested for evidence of non-additive (dominant or recessive) effects, SNP × SNP interaction effects and heterogeneity by sex or study among the 32 BMI-associated SNPs (Online Methods). We found no evidence for any such effects (all $P > 0.001$ and no significant results were seen after correcting for multiple testing) (**Supplementary Table 1** and **Supplementary Note**).

Impact of the 32 confirmed loci on BMI, obesity, body size and other metabolic traits

Together, the 32 confirmed BMI loci explained 1.45% of the inter-individual variation in BMI in the stage 2 samples, with the *FTO* SNP accounting for the largest proportion of the variance (0.34%) (**Table 1**). To estimate the cumulative effect of the 32 variants on BMI, we constructed a genetic susceptibility score that summed the number of BMI-increasing alleles weighted by the overall stage 2 effect sizes in the Atherosclerosis Risk in Communities (ARIC) study (n = 8,120), one of our largest population-based studies (Online Methods). For each unit increase in the genetic-susceptibility score, which is approximately equivalent to having one additional risk allele, BMI increased by 0.17 kg/m^2, the equivalent of a 435–551 g gain in body weight in adults of 160–180 cm in height. The difference in average BMI between individuals with a high genetic-susceptibility score (defined as having ≥38 BMI-increasing alleles, comprising 1.5% (n = 124) of the ARIC sample) and those with a low genetic-susceptibility score (defined as having ≤21 BMI-increasing alleles, comprising 2.2% (n = 175) of the ARIC sample) was 2.73 kg/m^2, equivalent to a 6.99–8.85 kg body weight difference in adults of 160–180 cm in height (**Fig. 2a**). Still, we note that the predictive value for obesity risk and BMI of the 32 variants combined was modest, although it was statistically significant (**Fig. 2b** and **Supplementary Fig. 4**). The area under the receiver-operating characteristic (ROC) curve for prediction of risk of obesity (BMI ≥ 30 kg/m^2) using a model including age, age^2 and sex only was 0.515 (P = 0.023 compared to the area under the curve (AUC) of 0.50), which increased to 0.575 ($P < 10^{-5}$) when the 32 confirmed SNPs were also included in the model (**Fig. 2b**). The area under the ROC curve for the model including the 32 SNPs only was 0.574 ($P < 10^{-5}$).

All 32 confirmed BMI-increasing alleles showed directionally consistent effects on the risk of being overweight (BMI ≥ 25 kg/m^2) or obese (BMI ≥ 30 kg/m^2) in the stage 2 samples, with 30 of 32 variants achieving at least nominally significant associations. The BMI-increasing alleles increased the odds of being overweight by 1.013- to 1.138-fold and the odds of being obese by 1.016- to 1.203-fold (**Supplementary Table 2**). In addition, 30 of the 32 loci also showed directionally consistent effects on the risk of extreme and early-onset obesity in a meta-analysis of seven case-control studies of adults and children (binomial sign test $P = 1.3 \times 10^{-7}$) (**Supplementary Table 3**). The BMI-increasing allele observed in adults also increased the BMI in children and adolescents with directionally consistent effects observed for 23 of the 32 SNPs (binomial sign test P = 0.01). Furthermore, in family-based studies, the BMI-increasing allele was over-transmitted to the obese offspring for 24 of the 32 SNPs (binomial sign test P = 0.004) (**Supplementary Table 3**). As these studies in extreme obesity cases, children and families were relatively small (with n ranging from 354 to 15,251 individuals) compared to the overall meta-analyses, their power was likely insufficient to confirm association for all 32 loci. Nevertheless, these results show that the effects are unlikely to reflect population stratification and that they extend to BMI differences throughout the life course.

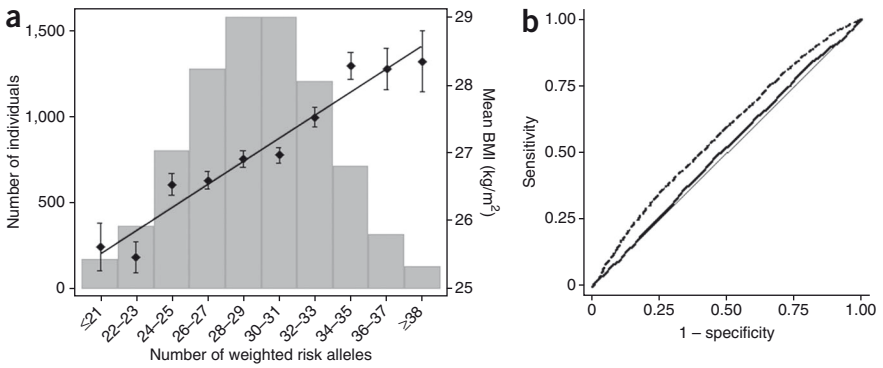


Figure 2 Combined impact of risk alleles on BMI and obesity. (**a**) Combined effect of risk alleles on average BMI in the population-based ARIC study ($n = 8,120$ individuals of European descent). For each individual, the number of 'best guess' replicated ($n = 32$) risk alleles from imputed data (0, 1 or 2) per SNP was weighted for its relative effect size estimated from the stage 2 data. Weighted risk alleles were summed for each individual, and the overall individual sum was rounded to the nearest integer to represent the individual's risk allele score (ranging from 16 to 44). Along the *x* axis, individuals in each risk allele category are shown (grouped as having ≤21 risk alleles and ≥38 risk alleles at the extremes), and the mean BMI (± s.e.m.) is plotted (*y* axis on right), with the line representing the regression of the mean BMI values across the risk-allele scores. The histogram (*y* axis on left) represents the number of individuals in each risk-score category. (**b**) The area under the ROC curve (AUC) of two different models predicting the risk of obesity (BMI ≥ 30 kg/m^2) in the 8,120 genotyped individuals of European descent in the ARIC study. Model 1, represented by the solid line, includes age, age^2 and sex (AUC = 0.515, $P = 0.023$ for difference from the null AUC = 0.50). Model 2, represented by the dashed line, includes age, age^2, sex and the 32 confirmed BMI SNPs (AUC = 0.575, $P < 10^{-5}$ for difference from the null AUC = 0.50). The difference between both AUCs is significant ($P < 10^{-4}$).

All BMI-increasing alleles were associated with increased body weight, as could be expected from the correlation between BMI and body weight (**Supplementary Table 2**). To confirm an effect of the loci on adiposity rather than general body size, we tested for association with body fat percentage, for which data was available in a subset of the stage 2 replication samples ($n = 5,359$ to $n = 28,425$) (**Supplementary Table 2**). The BMI-increasing allele showed directionally consistent effects on body fat percentage at 31 of the 32 confirmed loci (binomial sign test $P = 1.54 \times 10^{-8}$) (**Supplementary Table 2**).

We also examined the association of the BMI loci with metabolic traits (type 2 diabetes[18], fasting glucose, fasting insulin, indices of β-cell function (HOMA-B) and insulin resistance (HOMA-IR)[19], and blood lipid levels[20]) and with height (**Supplementary Tables 2** and **4**). Although many nominal associations were expected because of known correlations between BMI and most of these traits, and because of overlap in samples, several associations stood out as possible examples of pleiotropic effects of the BMI-associated variants. Particularly interesting is the variant in the *GIPR* locus, where the BMI-increasing allele is also associated with increased fasting glucose levels and lower 2-h glucose levels (**Supplementary Table 4**)[19,21]. The direction of the effect is opposite to what would be expected due to the correlation between obesity and glucose intolerance but is consistent with the suggested roles of *GIPR* in glucose and energy metabolism (see below)[22]. Three loci showed strong associations ($P < 10^{-4}$) with height (at *MC4R*, *RBJ-ADCY3-POMC* and *MTCH2-NDUFS3*). Because BMI is weakly correlated with height (and indeed, the BMI-associated variants as a group show no consistent effect on height), these associations are also suggestive of pleiotropy. Notably, analogous to the effects of severe mutations in *POMC* and *MC4R* on height and weight[23,24], the BMI-increasing alleles of the variants near these genes were associated with decreased (*POMC*) and increased (*MC4R*) height, respectively (**Supplementary Table 2**).

Potential functional roles and pathway analyses

Although associated variants typically implicate genomic regions rather than individual genes, we note that some of the 32 loci include candidate genes with established connections to obesity. Several of the ten previously identified loci are located in or near genes that encode neuronal regulators of appetite or energy balance, including *MC4R*[12,25], *BDNF*[26] and *SH2B1*[11,27]. Each of these genes has been tied to obesity, not only in animal models, but also by rare human variants that disrupt each of these genes and lead to severe obesity[24,28,29]. Using the automated literature search program Snipper (Online Methods), we identified various genes within the newly discovered loci with potential biological links to obesity susceptibility (**Supplementary Note**). Among the new loci, the location of rs713586 near *POMC* provides further support for a role of neuroendocrine circuits that regulate energy balance in susceptibility to obesity. *POMC* encodes several polypeptides, including α-MSH, a ligand of the *MC4R* gene product[30], and rare mutations in *POMC* also cause obesity in humans[23,29,31].

In contrast, the locus near *GIPR*, which encodes a receptor of gastric inhibitory polypeptide (GIP), suggests a role for peripheral biology in obesity. GIP, which is expressed in the K cell of the duodenum and intestine, is an incretin hormone that mediates incremental insulin secretion in response to oral intake of glucose. The variant associated with BMI is in strong LD ($r^2 = 0.83$) with a missense SNP in *GIPR* (rs1800437, p.Glu354Gln) that has recently been shown to influence glucose and insulin response to an oral glucose challenge[21]. Although no human phenotype is known to be caused by mutations in *GIPR*, mice with disruption of *Gipr* are resistant to diet-induced obesity[32]. The association of a variant in *GIPR* with BMI suggests that there may be a link between incretins, insulin secretion and body weight regulation in humans as well.

To systematically identify biological connections among the genes located near the 32 confirmed SNPs and to potentially identify new pathways associated with BMI, we performed pathway-based analyses using MAGENTA[33]. Specifically, we tested for enrichment of genetic associations to BMI in biological processes or molecular functions that contain at least one gene from the 32 confirmed BMI loci (Online Methods). Using annotations from the Kyoto Encyclopedia of Genes and Genomes (KEGG), Ingenuity, Protein Analysis Through Evolutionary Relationships (PANTHER) and Gene Ontology databases, we found evidence of enrichment for pathways involved in platelet-derived growth factor (PDGF) signaling (PANTHER, $P = 0.0008$, false discovery rate (FDR) = 0.0061), translation elongation (PANTHER, $P = 0.0008$, FDR = 0.0066), hormone or nuclear-hormone receptor binding (Gene Ontology, $P < 0.0005$, FDR < 0.0085), homeobox transcription (PANTHER, $P = 0.0001$, FDR = 0.011), regulation of cellular metabolism (Gene Ontology, $P = 0.0002$, FDR = 0.031), neurogenesis and neuron differentiation (Gene Ontology, $P < 0.0002$, FDR < 0.034), protein phosphorylation (PANTHER, $P = 0.0001$, FDR = 0.045)

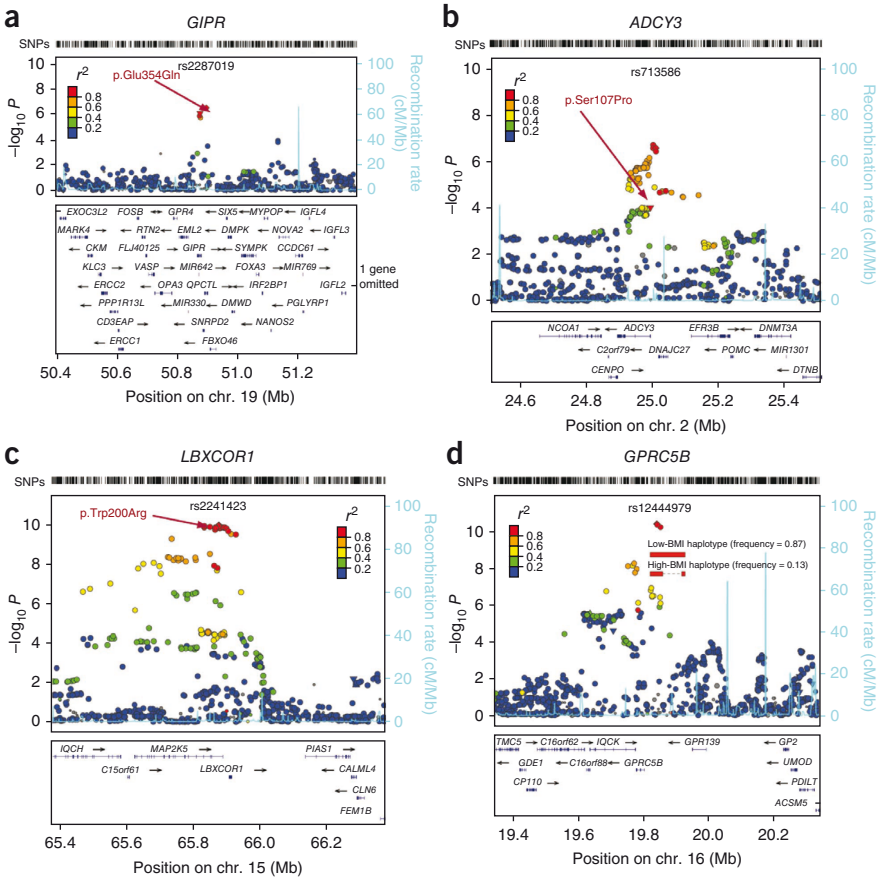


Figure 3 Regional plots of selected replicating BMI loci with missense and CNV variants. SNPs are plotted by position on the chromosome against association with BMI ($-\log_{10} P$). The SNP name shown on the plot was the most significant SNP after the stage 1 meta-analysis. Estimated recombination rates (from HapMap) are plotted in cyan to reflect the local LD structure. The SNPs surrounding the most significant SNP are color coded to reflect their LD with this SNP (taken from pairwise r^2 values from the HapMap CEU data). Genes, the position of exons and the direction of transcription from the UCSC genome browser are noted. Hashmarks represent SNP positions available in the meta-analysis. (**a**–**c**) Missense variants noted with their amino acid change for the gene listed above the plot. (**d**) Structural haplotypes and the BMI association signal in the *GPRC5B* region. A 21-kb deletion polymorphism was associated with four SNPs ($r^2 = 1.0$) that comprise the best haplogroup associating with BMI. Plots were generated using LocusZoom (see URLs).

and numerous other pathways related to growth, metabolism, immune and neuronal processes (Gene Ontology, $P < 0.002$, FDR < 0.046) (**Supplementary Table 5**).

Identifying possible functional variants

We used data from the 1000 Genomes Project and the HapMap Consortium to explore whether the 32 confirmed BMI SNPs were in LD ($r^2 \geq 0.75$) with common missense SNPs or copy number variants (CNVs) (Online Methods). Non-synonymous variants in LD with our signals were present in *BDNF*, *SLC39A8*, *FLJ35779-HMGCR*, *QPCTL-GIPR*, *MTCH2*, *ADCY3* and *LBXCOR1*. In addition, the rs7359397 signal was in LD with coding variants in several genes including *SH2B1*, *ATNX2L*, *APOB48R*, *SULT1A2* and *AC138894.2* (**Table 1**, **Fig. 3**, **Supplementary Table 6** and **Supplementary Fig. 2**). Furthermore, two SNPs tagged common CNVs. The first CNV has been previously identified[9] and is a 45-kb deletion near *NEGR1*. The second CNV is a 21-kb deletion that lies 50 kb upstream of *GPRC5B*; the deletion allele is tagged by the T allele of rs12444979 ($r^2 = 1$) (**Fig. 3**). Although the correlations with potentially functional variants do not prove that these variants are indeed causal, they provide first clues as to which genes and variants at these loci might be prioritized for fine mapping and functional follow up.

Because many of the 32 BMI loci harbor multiple genes, we examined whether gene expression quantitative trait loci (eQTL) analyses could also direct us to positional candidates. Gene expression data were available for human brain, lymphocyte, blood, subcutaneous and visceral adipose tissue, and liver[34–36] (Online Methods, **Table 1** and **Supplementary Table 7**). Significant *cis* associations, defined at the tissue-specific level, were observed between 14 BMI-associated alleles and expression levels (**Table 1** and **Supplementary Table 7**). In several instances, the BMI-associated SNP was the most significant SNP or explained a substantial proportion of the association with the most significant SNP for the gene transcript in conditional analyses (adjusted $P > 0.05$). These significant associations included *NEGR1*, *ZC3H4*, *TMEM160*, *MTCH2*, *NDUFS3*, *GTF3A*, *ADCY3*, *APOB48R*, *SH2B1*, *TUFM*, *GPRC5B*, *IQCK*, *SLC39A8*, *SULT1A1* and *SULT1A2* (**Table 1** and **Supplementary Table 7**), making these genes higher priority candidates within the associated loci. However, we note that some BMI-associated variants were correlated with the expression of multiple nearby genes, making it difficult to determine the most relevant gene.

Evidence for the existence of additional associated variants

Because the variants identified by this large study explain only 1.45% of the variance in BMI (2%–4% of genetic variance based on an estimated heritability of 40%–70%), we considered how much the explained phenotypic variance could be increased by including more SNPs at various degrees of significance in a polygene model using an independent validation set (Online Methods)[37]. We found that including SNPs associated with BMI at lower significance levels (up to $P > 0.05$) increased the explained phenotypic variance in BMI to 2.5%, or 4%–6% of the genetic variance (**Fig. 4a**). In a separate analysis, we estimated the total number of independent BMI-associated variants that are likely to exist with similar effect sizes as the 32 confirmed here (Online Methods)[38]. Based on the effect size and allele frequencies of the 32 replicated loci observed in stage 2 and the power to detect association in stage 1 and stage 2 combined, we estimated that there are 284 (95% CI 132–510) loci with similar effect sizes as those currently observed, which together would account for 4.5% (95% CI 3.1%–6.8%) of the phenotypic variation or 6%–11% of the genetic variation in BMI (based on an estimated heritability of 40%–70%) (**Supplementary Table 8**). In order to detect 95% of these loci, a sample size of approximately 730,000 subjects would be needed (**Fig. 4b**). This method does not account for the potential of loci of smaller effect than those identified here to explain even more of the variance and thus provides an estimated lower bound of explained variance. These two analyses strongly suggest that larger GWAS will

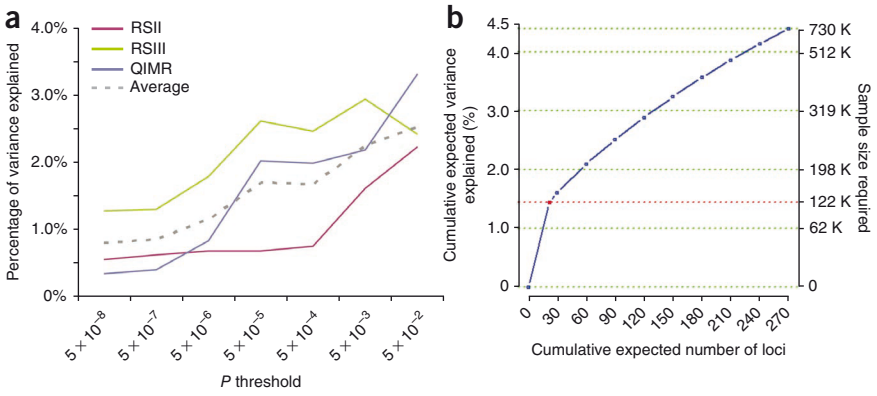


Figure 4 Phenotypic variance explained by common variants. (**a**) The variance explained is higher when SNPs not reaching genome-wide significance are included in the prediction model. The *y* axis represents the proportion of variance explained at different *P* value thresholds from the stage 1 meta-analysis. Results are given for three studies (Rotterdam Study II (RSII), Rotterdam Study III (RSIII), Queens Institute of Medical Research (QIMR)) which were not included in the meta-analysis, after exclusion of all samples from The Netherlands (for RSII and RSIII) and the United Kingdom (for QIMR) from the discovery analysis for this sub-analysis. The dotted line represents the weighted average of the explained variance of three validation sets. (**b**) Cumulative number of susceptibility loci expected to be discovered, including those we have already identified and others that have yet to be detected, by the expected percentage of phenotypic variation explained and the sample size required for a one-stage GWAS assuming a genomic control correction is used. The projections are based on loci that achieved a significance level of $P < 5 \times 10^{-8}$ in the joint analysis of stage 1 and stage 2 and the distribution of their effect sizes in stage 2. The dotted red line corresponds to the expected phenotypic variance explained by the 22 loci that are expected to be discovered in a one-stage GWAS using the sample size of stage 1 of this study.

continue to identify additional new associated loci but also indicate that even extremely large studies focusing on variants with allele frequencies above 5% will not account for a large fraction of the genetic contribution to BMI.

We examined whether selecting only a single variant from each locus for follow up led us to underestimate the fraction of phenotypic variation explained by the associated loci. To search for additional independent loci at each of the 32 associated BMI loci, we repeated our genome-wide association meta-analysis conditioning on the 32 confirmed SNPs. Using a significance threshold of $P = 5 \times 10^{-6}$ for SNPs at known loci, we identified one apparently independent signal at the *MC4R* locus; rs7227255 was associated with BMI ($P = 6.56 \times 10^{-7}$) even after conditioning for the most strongly associated variant near *MC4R* (rs571312) (**Fig. 5**). Notably, rs7227255 is in perfect LD ($r^2 = 1$) with a relatively rare *MC4R* missense variant (rs2229616, p.Val103Ile, minor allele frequency = 1.7%) that has been associated with BMI in two independent meta-analyses[39,40]. Furthermore, mutations at the *MC4R* locus are known to influence early-onset obesity[24,41], supporting the notion that allelic heterogeneity may be a frequent phenomenon in the genetic architecture of obesity.

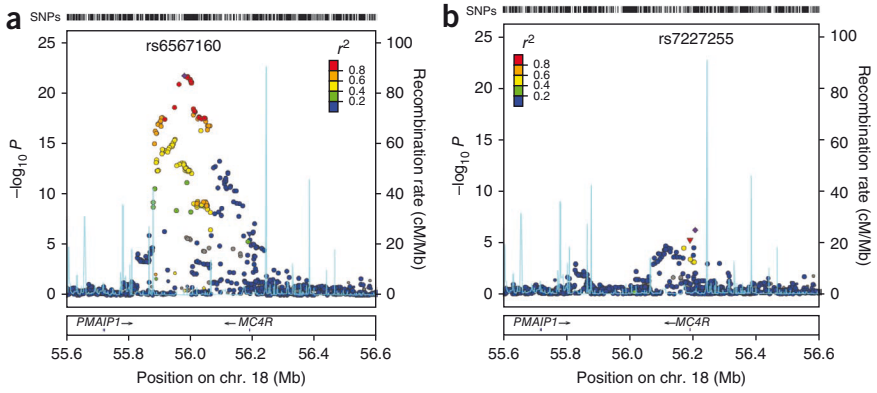


Figure 5 A second signal at the *MC4R* locus contributing to BMI. SNPs are plotted by position in a 1-Mb window of chromosome 18 against association with BMI ($-\log_{10} P$). (**a**) Plot highlighting the most significant SNP in the stage 1 meta-analysis. (**b**) Plot highlighting the most significant SNP after conditional analysis, where the model included the most strongly associated SNP as a covariate. Estimated recombination rates (from HapMap) are plotted in cyan to reflect the local LD structure. The SNPs surrounding the most significant SNP are color coded to reflect their LD with this SNP (taken from pairwise r^2 values from the HapMap CEU database). Genes, exons and the direction of transcription from the UCSC genome browser are noted. Hashmarks at the top of the figure represent the positions of SNPs in the meta-analysis. Regional plots were generated using LocusZoom.

DISCUSSION

Using a two-stage genome-wide association meta-analysis of up to 249,796 individuals of European descent, we identified 18 additional loci that are associated with BMI at genome-wide significance, bringing the total number of such loci to 32. We estimate that more than 250 common variant loci (that is, 284 predicted loci minus 32 confirmed loci) with effects on BMI similar to those described here remain to be discovered and that even larger numbers of loci with smaller effects remain to be identified. A substantial proportion of these loci should be identifiable through larger GWAS and/or by targeted follow up of the top signals selected from our stage 1 analysis. The latter approach is already being implemented through large-scale genotyping of samples informative for BMI using a custom array (the Metabochip) designed to support follow up of thousands of promising variants in hundreds of thousands of individuals.

The combined effect on BMI of the associated variants at the 32 loci is modest, and even when we try to account for as yet undiscovered variants with similar properties, we estimate that these common variant signals account for only 6%–11% of the genetic variation in BMI. There is a strong expectation that additional variance and biology will be explained using complementary approaches that capture variants not examined in the current study, such as lower frequency variants and short insertion-deletion polymorphisms. There is good reason to believe (based on our findings at *MC4R* and other loci, such as those at *POMC*, *BDNF* and *SH2B1*, which feature both common and rare variant associations) that a proportion of such low-frequency and rare causal variation will map to the loci already identified by GWAS.

A primary goal of human genetic discovery is to improve understanding of the biology

of conditions such as obesity[42]. One particularly noteworthy finding in this regard is the association between BMI and common variants near *GIPR*, which may indicate a causal contribution of variation in postprandial insulin secretion in the development of obesity. In most instances, the loci identified by the present study harbor few, if any, annotated genes with clear connections to the biology of weight regulation. This reflects our still limited understanding of the biology of BMI and obesity-related traits and is in striking contrast with the results from equivalent studies of certain other traits (such as autoimmune diseases or lipid levels). Thus, these results suggest that much of the biology that underlies obesity remains to be uncovered and that GWAS may provide an important entry point for investigation. In particular, further examination of the associated loci through a combination of resequencing and fine mapping to find causal variants and genomic and experimental studies designed to assign function could uncover new insights into the biology of obesity.

In conclusion, we performed GWAS in large samples to identify numerous genetic loci associated with variation in BMI, a common measure of obesity. Because current lifestyle interventions are largely ineffective in addressing the challenges of growing obesity[43,44], new insights into the biology of obesity are critically needed to guide the development and application of future therapies and interventions.

URLs. LocusZoom, http://csg.sph.umich.edu/locuszoom; METAL, http://www.sph.umich.edu/csg/abecasis/Metal/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

A full list of acknowledgments appears in the **Supplementary Note**.

Funding was provided by Academy of Finland (10404, 77299, 104781, 114382, 117797, 120315, 121584, 124243, 126775, 126925, 127437, 129255, 129269, 129306, 129494, 129680, 130326, 209072, 210595, 213225, 213506 and 216374); ADA Mentor-Based Postdoctoral Fellowship; Amgen; Agency for Science, Technology and Research of Singapore (A*STAR); ALF/LUA research grant in Gothenburg; Althingi (the Icelandic Parliament); AstraZeneca; Augustinus Foundation; Australian National Health and Medical Research Council (241944, 389875, 389891, 389892, 389938, 442915, 442981, 496739, 496688, 552485 and 613672); Australian Research Council (ARC grant DP0770096); Becket Foundation; Biocenter (Finland); Biomedicum Helsinki Foundation; Boston Obesity Nutrition Research Center; British Diabetes Association (1192); British Heart Foundation (97020; PG/02/128); Busselton Population Medical Research Foundation; Cambridge Institute for Medical Research; Cambridge National Institute of Health Research (NIHR) Comprehensive Biomedical Research Centre; CamStrad (UK); Cancer Research UK; Centre for Medical Systems Biology (The Netherlands); Centre for Neurogenomics and Cognitive Research (The Netherlands); Chief Scientist Office of the Scottish Government; Contrat Plan Etat Région (France); Danish Centre for Health Technology Assessment; Danish Diabetes Association; Danish Heart Foundation; Danish Pharmaceutical Association; Danish Research Council; Deutsche Forschungsgemeinschaft (DFG; HE 1446/4-1); Department of Health (UK); Diabetes UK; Diabetes and Inflammation Laboratory; Donald W. Reynolds Foundation; Dresden University of Technology Funding Grant; Emil and Vera Cornell Foundation; Erasmus Medical Center (Rotterdam); Erasmus University (Rotterdam); European Commission (DG XII; QLG1-CT-2000-01643, QLG2-CT-2002-01254, LSHC-CT-2005, LSHG-CT-2006-018947, LSHG-CT-2004-518153, LSH-2006-037593, LSHM-CT-2007-037273, HEALTH-F2-2008-ENGAGE, HEALTH-F4-2007-201413, HEALTH-F4-2007-201550, FP7/2007-2013, 205419, 212111, 245536, SOC 95201408 05F02 and WLRT-2001-01254); Federal Ministry of Education and Research (Germany) (01AK803, 01EA9401, 01GI0823, 01GI0826, 01GP0209, 01GP0259, 01GS0820, 01GS0823, 01GS0824, 01GS0825, 01GS0830, 01GS0831, 01IG07015, 01KU0903, 01ZZ9603, 01ZZ0103, 01ZZ0403 and 03ZIK012); Federal State of Mecklenburg-West Pomerania; European Social Fund; Eve Appeal; Finnish Diabetes Research Foundation; Finnish Foundation for Cardiovascular Research; Finnish Foundation for Pediatric Research, Finnish Medical Society; Finska Läkaresällskapet, Päivikki and Sakari Sohlberg Foundation, Folkhälsan Research Foundation; Fond Européen pour le Développement Régional (France); Fondation LeDucq (Paris, France); Foundation for Life and Health in Finland; Foundation for Strategic Research (Sweden); Genetic Association Information Network; German Research Council (KFO-152); German National Genome Research Net 'NGFNplus' (FKZ 01GS0823); German Research Center for Environmental Health; Giorgi-Cavaglieri Foundation; GlaxoSmithKline; Göteborg Medical Society; Great Wine Estates Auctions; Gyllenberg Foundation; Health Care Centers in Vasa, Närpes and Korsholm; Healthway, Western Australia; Helmholtz Center Munich; Helsinki University Central Hospital, Hjartavernd (the Icelandic Heart Association); INSERM (France); Ib Henriksen Foundation; Interdisziplinäres Zentrum für Klinische Forschung (IZKF) (B27); Jalmari and Rauha Ahokas Foundation; Juho Vainio Foundation; Juvenile Diabetes Research Foundation International (JDRF); Karolinska Institute; Knut and Alice Wallenberg Foundation; Leenaards Foundation; Lundbeck Foundation Centre of Applied Medical Genomics for Personalized Disease Prediction, Prevention and Care (LUCAMP); Lundberg Foundation; Marie Curie Intra-European Fellowship; Medical Research Council (UK) (G0000649, G0000934, G9521010D, G0500539, G0600331 and G0601261, PrevMetSyn); Ministry of Cultural Affairs and Social Ministry of the Federal State of Mecklenburg-West Pomerania; Ministry for Health, Welfare and Sports (The Netherlands); Ministry of Education (Finland); Ministry of Education, Culture and Science (The Netherlands); Ministry of Internal Affairs and Health (Denmark); Ministry of Science, Education and Sport of the Republic of Croatia (216-1080315-0302); Ministry of Science, Research and the Arts Baden-Württemberg; Montreal Heart Institute Foundation; Municipal Health Care Center and Hospital in Jakobstad; Municipality of Rotterdam; Närpes Health Care Foundation; National Cancer Institute; National Health and Medical Research Council of Australia; National Institute for Health Research Cambridge Biomedical Research Centre; National Institute for Health Research Oxford Biomedical Research Centre; National Institute for Health Research comprehensive Biomedical Research Centre; US National Institutes of Health (263-MA-410953, AA07535, AA10248, AA014041, AA13320, AA13321, AA13326, CA047988, CA65725, CA87969, CA49449, CA67262, CA50385, DA12854, DK58845, DK46200, DK062370, DK063491, DK072193, HG002651, HL084729, HHSN268200625226C, HL71981, K23-DK080145, K99-HL094535, M01-RR00425, MH084698, N01-AG12100, NO1-AG12109, N01-HC15103, N01-HC25195, N01-HC35129, N01-HC45133, N01-HC55015, N01-HC55016, N01-HC55018, N01-HC55019, N01-HC55020, N01-N01HC-55021, N01-HC55022, N01-HC55222, N01-HC75150, N01-HC85079, N01-HC85080, N01-HG-65403, N01-HC85081, N01-HC85082, N01-HC85083, N01-HC85084, N01-HC85085, N01-HC85086, N02-HL64278, P30-DK072488, R01-AG031890, R01-DK073490, R01-DK075787, R01DK068336, R01DK075681, R01-HL59367, R01-HL086694, R01-HL087641, R01-HL087647, R01-HL087652, R01-HL087676, R01-HL087679, R01-HL087700, R01-HL088119, R01-MH59160, R01-MH59565, R01-MH59566, R01-MH59571, R01-MH59586, R01-MH59587, R01-MH59588, R01-MH60870, R01-MH60879, R01-MH61675, R01-MH63706, R01-MH67257, R01-MH79469, R01-MH79470, R01-MH81800, RL1-MH083268, UO1-CA098233, U01-DK062418, U01-GM074518, U01-HG004402, U01-HG004399, U01-HL72515, U01-HL080295, U01-HL084756, U54-RR020278, T32-HG00040, UL1-RR025005 and Z01-HG000024); National Alliance for Research on Schizophrenia and Depression (NARSAD); Netherlands Genomics Initiative/Netherlands Consortium for Healthy Aging (050-060-810); Netherlands Organisation for Scientific Research (NWO) (904-61-090, 904-61-193, 480-04-004, 400-05-717, SPI 56-464-1419, 175.010.2005.011 and 911-03-012); Nord-Trøndelag County Council; Nordic Center of Excellence in Disease Genetics; Novo Nordisk Foundation; Norwegian Institute of Public Health; Ollqvist Foundation; Oxford NIHR Biomedical Research Centre; Organization for the Health Research and Development (10-000-1002); Paavo Nurmi Foundation; Paul Michael Donovan Charitable Foundation; Perklén Foundation; Petrus and Augusta Hedlunds Foundation; Pew Scholar for the Biomedical Sciences; Public Health and Risk Assessment, Health and Consumer Protection (2004310); Research Foundation of Copenhagen County; Research Institute for Diseases in the Elderly (014-93-015; RIDE2); Robert Dawson Evans Endowment; Royal Society (UK); Royal Swedish Academy of Science; Sahlgrenska Center for Cardiovascular and Metabolic Research (CMR, no. A305: 188); Siemens Healthcare, Erlangen, Germany; Sigrid Juselius Foundation; Signe and Ane Gyllenberg Foundation; Science Funding programme (UK); Social Insurance Institution of Finland; Söderberg's Foundation; South Tyrol Ministry of Health; South Tyrolean Sparkasse Foundation; State of Bavaria; Stockholm County Council (560183); Susan G. Komen Breast Cancer Foundation; Swedish Cancer Society; Swedish Cultural Foundation in Finland; Swedish Foundation for Strategic Research; Swedish Heart-Lung Foundation; Swedish Medical Research Council (8691, K2007-66X-20270-01-3, K2010-54X-09894-19-3, K2010-54X-09894-19-3 and 2006-3832); Swedish Research Council; Swedish Society of Medicine; Swiss National Science Foundation

(33CSCO-122661, 310000-112552 and 3100AO-116323/1); Torsten and Ragnar Söderberg's Foundation; Université Henri Poincaré-Nancy 1, Région Lorraine, Communauté Urbaine du Grand Nancy; University Hospital Medical funds to Tampere; University Hospital Oulu, Finland; University of Oulu, Finland (75617); Västra Götaland Foundation; Walter E. Nichols, M.D., and Eleanor Nichols endowments; Wellcome Trust (068545, 072960, 075491, 076113, 077016, 079557, 079895, 081682, 083270, 085301 and 086596); Western Australian DNA Bank; Western Australian Genetic Epidemiology Resource; and Yrjö Jahnsson Foundation.

AUTHOR CONTRIBUTIONS

A full list of author contributions appears in the **Supplementary Note**.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Anonymous. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults–the evidence report. National Institutes of Health. *Obes. Res.* **6** Suppl 2, 51S–209S (1998); erratum *Obes. Res.* **6**, 464 (1998); comment *Obes. Res.* **6**, 461–462 (1998).
2. Lewis, C.E. *et al.* Mortality, health outcomes, and body mass index in the overweight range: a science advisory from the American Heart Association. *Circulation* **119**, 3263–3271 (2009).
3. Stunkard, A.J., Foch, T.T. & Hrubec, Z. A twin study of human obesity. *J. Am. Med. Assoc.* **256**, 51–54 (1986).
4. Maes, H.H., Neale, M.C. & Eaves, L.J. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.* **27**, 325–351 (1997).
5. Taylor, A.E. *et al.* Comparison of the associations of body mass index and measures of central adiposity and fat mass with coronary heart disease, diabetes, and all-cause mortality: a study using data from 4 UK cohorts. *Am. J. Clin. Nutr.* **91**, 547–556 (2010).
6. Frayling, T.M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
7. Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLoS Genet.* **3**, e115 (2007).
8. Loos, R.J. *et al.* Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nat. Genet.* **40**, 768–775 (2008).
9. Willer, C.J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).
10. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* **41**, 18–24 (2009).
11. Ren, D. *et al.* Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. *J. Clin. Invest.* **117**, 397–406 (2007).
12. Huszar, D. *et al.* Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88**, 131–141 (1997).
13. O'Rahilly, S. & Farooqi, I.S. Human obesity as a heritable disorder of the central control of energy balance. *Int. J. Obes. (Lond)* **32** Suppl 7, S55–S61 (2008).
14. Lindgren, C.M. *et al.* Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* **5**, e1000508 (2009).
15. Heard-Costa, N.L. *et al. NRXN3* is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS Genet.* **5**, e1000539 (2009).
16. Meyre, D. *et al.* Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* **41**, 157–159 (2009).
17. Scherag, A. *et al.* Two new loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. *PLoS Genet.* **6**, e1000916 (2010).
18. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638–645 (2008).
19. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
20. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
21. Saxena, R. *et al.* Genetic variation in *GIPR* influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42**, 142–148 (2010).
22. McIntosh, C.H., Widenmaier, S. & Kim, S.J. Glucose-dependent insulinotropic polypeptide (Gastric Inhibitory Polypeptide, GIP). *Vitam. Horm.* **80**, 409–471 (2009).
23. Farooqi, I.S. *et al.* Heterozygosity for a POMC-null mutation and increased obesity risk in humans. *Diabetes* **55**, 2549–2553 (2006).
24. Farooqi, I.S. *et al.* Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *N. Engl. J. Med.* **348**, 1085–1095 (2003).
25. Marsh, D.J. *et al.* Response of melanocortin-4 receptor-deficient mice to anorectic and orexigenic peptides. *Nat. Genet.* **21**, 119–122 (1999).
26. Unger, T.J., Calderon, G.A., Bradley, L.C., Sena-Esteves, M. & Rios, M. Selective deletion of Bdnf in the ventromedial and dorsomedial hypothalamus of adult mice results in hyperphagic behavior and obesity. *J. Neurosci.* **27**, 14265–14274 (2007).
27. Li, Z., Zhou, Y., Carter-Su, C., Myers, M.G. Jr. & Rui, L. SH2B1 enhances leptin signaling by both Janus kinase 2 Tyr813 phosphorylation-dependent and -independent mechanisms. *Mol. Endocrinol.* **21**, 2270–2281 (2007).
28. Gray, J. *et al.* Hyperphagia, severe obesity, impaired cognitive function, and hyperactivity associated with functional loss of one copy of the brain-derived neurotrophic factor (BDNF) gene. *Diabetes* **55**, 3366–3371 (2006).
29. Bochukova, E.G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666–670 (2010).
30. Coll, A.P. & Loraine Tung, Y.C. Pro-opiomelanocortin (POMC)-derived peptides and the regulation of energy homeostasis. *Mol. Cell. Endocrinol.* **300**, 147–151 (2009).
31. Krude, H. *et al.* Obesity due to proopiomelanocortin deficiency: three new cases and treatment trials with thyroid hormone and ACTH4–10. *J. Clin. Endocrinol. Metab.* **88**, 4633–4640 (2003).
32. Miyawaki, K. *et al.* Inhibition of gastric inhibitory polypeptide signaling prevents obesity. *Nat. Med.* **8**, 738–742 (2002).
33. Segrè, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).
34. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
35. Myers, A.J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39**, 1494–1499 (2007).
36. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 (2007).
37. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
38. Park, J.-H. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
39. Young, E.H. *et al.* The V103I polymorphism of the *MC4R* gene and obesity: population based studies and meta-analysis of 29 563 individuals. *Int. J. Obes. (Lond)* **31**, 1437–1441 (2007).
40. Stutzmann, F. *et al.* Non-synonymous polymorphisms in melanocortin-4 receptor protect against obesity: the two facets of a Janus obesity gene. *Hum. Mol. Genet.* **16**, 1837–1844 (2007).
41. Yeo, G.S. *et al.* Mutations in the human melanocortin-4 receptor gene associated with severe familial obesity disrupts receptor function through multiple molecular mechanisms. *Hum. Mol. Genet.* **12**, 561–574 (2003).
42. Hirschhorn, J.N. Genomewide association studies—illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).
43. Lemmens, V.E., Oenema, A., Klepp, K.I., Henriksen, H.B. & Brug, J. A systematic review of the evidence regarding efficacy of obesity prevention interventions among adults. *Obes. Rev.* **9**, 446–455 (2008).
44. Anderson, J.W., Konz, E.C., Frederich, R.C. & Wood, C.L. Long-term weight-loss maintenance: a meta-analysis of US studies. *Am. J. Clin. Nutr.* **74**, 579–584 (2001).

Elizabeth K Speliotes[1,2,250], Cristen J Willer[3,250], Sonja I Berndt[4,250], Keri L Monda[5,250], Gudmar Thorleifsson[6,250], Anne U Jackson[3], Hana Lango Allen[7], Cecilia M Lindgren[8,9], Jian'an Luan[10], Reedik Mägi[8], Joshua C Randall[8], Sailaja Vedantam[1,11], Thomas W Winkler[12], Lu Qi[13,14], Tsegaselassie Workalemahu[13], Iris M Heid[12,15], Valgerdur Steinthorsdottir[6], Heather M Stringham[3], Michael N Weedon[7], Eleanor Wheeler[16], Andrew R Wood[7], Teresa Ferreira[8], Robert J Weyant[3], Ayellet V Segrè[17–19], Karol Estrada[20–22], Liming Liang[23,24], James Nemesh[18],

Ju-Hyun Park[4], Stefan Gustafsson[25], Tuomas O Kilpeläinen[10], Jian Yang[26], Nabila Bouatia-Naji[27,28], Tõnu Esko[29–31], Mary F Feitosa[32], Zoltán Kutalik[33,34], Massimo Mangino[35], Soumya Raychaudhuri[18,36], Andre Scherag[37], Albert Vernon Smith[38,39], Ryan Welch[3], Jing Hua Zhao[10], Katja K Aben[40], Devin M Absher[41], Najaf Amin[20], Anna L Dixon[42], Eva Fisher[43], Nicole L Glazer[44,45], Michael E Goddard[46,47], Nancy L Heard-Costa[48], Volker Hoesel[49], Jouke-Jan Hottenga[50], Åsa Johansson[51,52], Toby Johnson[33,34,53,54], Shamika Ketkar[32], Claudia Lamina[15,55], Shengxu Li[10], Miriam F Moffatt[56], Richard H Myers[57], Narisu Narisu[58], John R B Perry[7], Marjolein J Peters[21,22], Michael Preuss[59], Samuli Ripatti[60,61], Fernando Rivadeneira[20–22], Camilla Sandholt[62], Laura J Scott[3], Nicholas J Timpson[63], Jonathan P Tyrer[64], Sophie van Wingerden[20], Richard M Watanabe[65,66], Charles C White[67], Fredrik Wiklund[25], Christina Barlassina[68], Daniel I Chasman[69,70], Matthew N Cooper[71], John-Olov Jansson[72], Robert W Lawrence[71], Niina Pellikka[60,61], Inga Prokopenko[8,9], Jianxin Shi[4], Elisabeth Thiering[15], Helene Alavere[29], Maria T S Alibrandi[73], Peter Almgren[74], Alice M Arnold[75,76], Thor Aspelund[38,39], Larry D Atwood[48], Beverley Balkau[77,78], Anthony J Balmforth[79], Amanda J Bennett[9], Yoav Ben-Shlomo[80], Richard N Bergman[66], Sven Bergmann[33,34], Heike Biebermann[81], Alexandra I F Blakemore[82], Tanja Boes[37], Lori L Bonnycastle[58], Stefan R Bornstein[83], Morris J Brown[84], Thomas A Buchanan[66,85], Fabio Busonero[86], Harry Campbell[87], Francesco P Cappuccio[88], Christine Cavalcanti-Proença[27,28], Yii-Der Ida Chen[89], Chih-Mei Chen[15], Peter S Chines[58], Robert Clarke[90], Lachlan Coin[91], John Connell[92], Ian N M Day[63], Martin den Heijer[93,94], Jubao Duan[95], Shah Ebrahim[96,97], Paul Elliott[91,98], Roberto Elosua[99], Gudny Eiriksdottir[38], Michael R Erdos[58], Johan G Eriksson[100–104], Maurizio F Facheris[105,106], Stephan B Felix[107], Pamela Fischer-Posovszky[108], Aaron R Folsom[109], Nele Friedrich[110], Nelson B Freimer[111], Mao Fu[112], Stefan Gaget[27,28], Pablo V Gejman[95], Eco J C Geus[50], Christian Gieger[15], Anette P Gjesing[62], Anuj Goel[8,113], Philippe Goyette[114], Harald Grallert[15], Jürgen Gräßler[115], Danielle M Greenawalt[116], Christopher J Groves[9], Vilmundur Gudnason[38,39], Candace Guiducci[1], Anna-Liisa Hartikainen[117], Neelam Hassanali[9], Alistair S Hall[79], Aki S Havulinna[118], Caroline Hayward[119], Andrew C Heath[120], Christian Hengstenberg[121,122], Andrew A Hicks[105], Anke Hinney[123], Albert Hofman[20,22], Georg Homuth[124], Jennie Hui[71,125,126], Wilmar Igl[51], Carlos Iribarren[127,128], Bo Isomaa[103,129], Kevin B Jacobs[130], Ivonne Jarick[131], Elizabeth Jewell[3], Ulrich John[132], Torben Jørgensen[133,134], Pekka Jousilahti[118], Antti Jula[135], Marika Kaakinen[136,137], Eero Kajantie[101,138], Lee M Kaplan[2,70,139], Sekar Kathiresan[17,18,140–142], Johannes Kettunen[60,61], Leena Kinnunen[143], Joshua W Knowles[144], Ivana Kolcic[145], Inke R König[59], Seppo Koskinen[118], Peter Kovacs[146], Johanna Kuusisto[147], Peter Kraft[23,24], Kirsti Kvaløy[148], Jaana Laitinen[149], Olivier Lantieri[150], Chiara Lanzani[73], Lenore J Launer[151], Cecile Lecoeur[27,28], Terho Lehtimäki[152], Guillaume Lettre[114,153], Jianjun Liu[154], Marja-Liisa Lokki[155], Mattias Lorentzon[156], Robert N Luben[157], Barbara Ludwig[83], MAGIC[158], Paolo Manunta[73], Diana Marek[33,34], Michel Marre[159,160], Nicholas G Martin[161], Wendy L McArdle[162], Anne McCarthy[163], Barbara McKnight[75], Thomas Meitinger[164,165], Olle Melander[166], David Meyre[27,28], Kristian Midthjell[148], Grant W Montgomery[167], Mario A Morken[58], Andrew P Morris[8], Rosanda Mulic[168], Julius S Ngwa[67], Mari Nelis[29–31], Matt J Neville[9], Dale R Nyholt[169], Christopher J O'Donnell[141,170], Stephen O'Rahilly[171], Ken K Ong[10], Ben Oostra[172], Guillaume Paré[173], Alex N Parker[174], Markus Perola[60,61], Irene Pichler[105], Kirsi H Pietiläinen[175,176], Carl G P Platou[148,177], Ozren Polasek[145,178], Anneli Pouta[117,179], Suzanne Rafelt[180], Olli Raitakari[181,182], Nigel W Rayner[8,9], Martin Ridderstråle[166], Winfried Rief[183], Aimo Ruokonen[184], Neil R Robertson[8,9], Peter Rzehak[15,185], Veikko Salomaa[118], Alan R Sanders[95], Manjinder S Sandhu[10,16,157], Serena Sanna[86], Jouko Saramies[186], Markku J Savolainen[187], Susann Scherag[123], Sabine Schipf[110,188], Stefan Schreiber[189], Heribert Schunkert[190], Kaisa Silander[60,61], Juha Sinisalo[191], David S Siscovick[45,192], Jan H Smit[193], Nicole Soranzo[16,35], Ulla Sovio[91], Jonathan Stephens[194,195], Ida Surakka[60,61], Amy J Swift[58], Mari-Liis Tammesoo[29], Jean-Claude Tardif[114,153], Maris Teder-Laving[30,31], Tanya M Teslovich[3], John R Thompson[196,197], Brian Thomson[1], Anke Tönjes[198,199], Tiinamaija Tuomi[103,200,201], Joyce B J van Meurs[20–22], Gert-Jan van Ommen[202,203], Vincent Vatin[27,28], Jorma Viikari[204], Sophie Visvikis-Siest[205], Veronique Vitart[119], Carla I G Vogel[123], Benjamin F Voight[17–19], Lindsay L Waite[41], Henri Wallaschofski[110], G Bragi Walters[6], Elisabeth Widen[60], Susanna Wiegand[81], Sarah H Wild[87], Gonneke Willemsen[50], Daniel R Witte[206], Jacqueline C Witteman[20,22], Jianfeng Xu[207], Qunyuan Zhang[32], Lina Zgaga[145], Andreas Ziegler[59], Paavo Zitting[208], John P Beilby[125,126,209], I Sadaf Farooqi[171], Johannes Hebebrand[123], Heikki V Huikuri[210], Alan L James[126,211], Mika Kähönen[212], Douglas F Levinson[213], Fabio Macciardi[68,214], Markku S Nieminen[191], Claes Ohlsson[156], Lyle J Palmer[71,126], Paul M Ridker[69,70], Michael Stumvoll[198,215], Jacques S Beckmann[33,216], Heiner Boeing[43], Eric Boerwinkle[217], Dorret I Boomsma[50], Mark J Caulfield[54], Stephen J Chanock[4], Francis S Collins[58],

L Adrienne Cupples[67], George Davey Smith[63], Jeanette Erdmann[190], Philippe Froguel[27,28,82], Henrik Grönberg[25], Ulf Gyllensten[51], Per Hall[25], Torben Hansen[62,218], Tamara B Harris[151], Andrew T Hattersley[7], Richard B Hayes[219], Joachim Heinrich[15], Frank B Hu[13,14,23], Kristian Hveem[148], Thomas Illig[15], Marjo-Riitta Jarvelin[91,136,137,179], Jaakko Kaprio[60,175,220], Fredrik Karpe[9,221], Kay-Tee Khaw[157], Lambertus A Kiemeney[40,93,222], Heiko Krude[81], Markku Laakso[147], Debbie A Lawlor[63], Andres Metspalu[29–31], Patricia B Munroe[54], Willem H Ouwehand[16,194,195], Oluf Pedersen[62,223,224], Brenda W Penninx[193,225,226], Annette Peters[15], Peter P Pramstaller[105,106,227], Thomas Quertermous[144], Thomas Reinehr[228], Aila Rissanen[176], Igor Rudan[87,168], Nilesh J Samani[180,196], Peter E H Schwarz[229], Alan R Shuldiner[112,230], Timothy D Spector[35], Jaakko Tuomilehto[143,231,232], Manuela Uda[86], André Uitterlinden[20–22], Timo T Valle[143], Martin Wabitsch[108], Gérard Waeber[233], Nicholas J Wareham[10], Hugh Watkins[8,113], on behalf of Procardis Consortium, James F Wilson[87], Alan F Wright[119], M Carola Zillikens[21,22], Nilanjan Chatterjee[4], Steven A McCarroll[17–19], Shaun Purcell[17,234,235], Eric E Schadt[236,237], Peter M Visscher[26], Themistocles L Assimes[144], Ingrid B Borecki[32,238], Panos Deloukas[16], Caroline S Fox[239], Leif C Groop[74], Talin Haritunians[89], David J Hunter[13,14,23], Robert C Kaplan[240], Karen L Mohlke[241], Jeffrey R O'Connell[112], Leena Peltonen[16,60,61,234,242], David Schlessinger[243], David P Strachan[244], Cornelia M van Duijn[20,22], H-Erich Wichmann[15,185,245], Timothy M Frayling[7], Unnur Thorsteinsdottir[6,246], Gonçalo R Abecasis[3], Inês Barroso[16,247], Michael Boehnke[3,250], Kari Stefansson[6,246,250], Kari E North[5,248,250], Mark I McCarthy[8,9,221,250], Joel N Hirschhorn[1,11,249,250], Erik Ingelsson[25,250] & Ruth J F Loos[10,250]

[1]Metabolism Initiative and Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [2]Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts, USA. [3]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA. [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA. [5]Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [6]deCODE Genetics, Reykjavik, Iceland. [7]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter, UK. [8]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [9]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK. [10]Medical Research Council (MRC) Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. [11]Divisions of Genetics and Endocrinology and Program in Genomics, Children's Hospital, Boston, Massachusetts, USA. [12]Regensburg University Medical Center, Department of Epidemiology and Preventive Medicine, Regensburg, Germany. [13]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA. [14]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [15]Institute of Epidemiology, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany. [16]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [17]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [18]Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [19]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA. [20]Department of Epidemiology, Erasmus Medical Center (MC), Rotterdam, The Netherlands. [21]Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. [22]Netherlands Genomics Initiative (NGI)-sponsored Netherlands Consortium for Healthy Aging (NCHA), Rotterdam, The Netherlands. [23]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [24]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [25]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [26]Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, Queensland, Australia. [27]Centre National de la Recherche Scientifique (CNRS) UMR8199-IBL-Institut Pasteur de Lille, Lille, France. [28]University Lille Nord de France, Lille, France. [29]Estonian Genome Center, University of Tartu, Tartu, Estonia. [30]Estonian Biocenter, Tartu, Estonia. [31]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia. [32]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA. [33]Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland. [34]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [35]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [36]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [37]Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Essen, Germany. [38]Icelandic Heart Association, Kopavogur, Iceland. [39]University of Iceland, Reykjavik, Iceland. [40]Comprehensive Cancer Center East, Nijmegen, The Netherlands. [41]Hudson Alpha Institute for Biotechnology, Huntsville, Alabama, USA. [42]Department of Pharmacy and Pharmacology, University of Bath, Bath, UK. [43]Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany. [44]Department of Medicine, University of Washington, Seattle, Washington, USA. [45]Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, USA. [46]University of Melbourne, Parkville, Australia. [47]Department of Primary Industries, Melbourne, Victoria, Australia. [48]Department of Neurology, Boston University School of Medicine, Boston, Massachusetts, USA. [49]Technical University Munich, Chair of Biomathematics, Garching, Germany. [50]Department of Biological Psychology, Vrije Universiteit (VU) University Amsterdam, Amsterdam, The Netherlands. [51]Department of Genetics and Pathology, Rudbeck Laboratory, University of Uppsala, Uppsala, Sweden. [52]Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [53]Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary, University of London, London, UK. [54]Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London, UK. [55]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria. [56]National Heart and Lung Institute, Imperial College London, London, UK. [57]Department of Neurology, Boston University School of Medicine, Boston, Massachusetts, USA. [58]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. [59]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany. [60]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [61]National Institute for Health and Welfare, Department of Chronic Disease Prevention, Unit of Public Health Genomics, Helsinki, Finland. [62]Hagedorn Research Institute, Gentofte, Denmark. [63]MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine, Oakfield House, Bristol, UK. [64]Department of Oncology, University of Cambridge, Cambridge, UK. [65]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [66]Department of Physiology and Biophysics, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [67]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA. [68]University of Milan, Department of Medicine, Surgery and Dentistry, Milano, Italy. [69]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [70]Harvard Medical School, Boston, Massachusetts, USA. [71]Centre for Genetic Epidemiology and Biostatistics, University of Western Australia, Crawley, Western Australia, Australia. [72]Department of Physiology, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [73]University Vita-Salute San Raffaele, Division of Nephrology and Dialysis, Milan, Italy. [74]Lund University Diabetes Centre, Department of Clinical Sciences, Lund University, Malmö, Sweden. [75]Department of Biostatistics, University of Washington, Seattle, Washington, USA. [76]Collaborative Health Studies Coordinating Center, Seattle, Washington, USA. [77]INSERM Centre de recherche en Epidémiologie et Santé des Populations (CESP) Centre for Research in Epidemiology and Public Health U1018, Villejuif, France. [78]University Paris Sud 11, Unité Mixte de Recherche en Santé (UMRS) 1018, Villejuif, France.

[79]Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, UK. [80]Department of Social Medicine, University of Bristol, Bristol, UK. [81]Institute of Experimental Paediatric Endocrinology, Charité Universitätsmedizin Berlin, Berlin, Germany. [82]Department of Genomics of Common Disease, School of Public Health, Imperial College London, London, UK. [83]Department of Medicine III, University of Dresden, Dresden, Germany. [84]Clinical Pharmacology Unit, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [85]Division of Endocrinology, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [86]Istituto di Neurogenetica e Neurofarmacologia del Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy. [87]Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh, Scotland, UK. [88]University of Warwick, Warwick Medical School, Coventry, UK. [89]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [90]Clinical Trial Service Unit, Oxford, UK. [91]Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, London, UK. [92]University of Dundee, Ninewells Hospital and Medical School, Dundee, UK. [93]Department of Epidemiology, Biostatistics and HTA, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [94]Department of Endocrinology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [95]Northshore University Healthsystem, Evanston, Illinois, USA. [96]The London School of Hygiene and Tropical Medicine, London, UK. [97]South Asia Network for Chronic Disease, New Dehli, India. [98]MRC-Health Protection Agency (HPA) Centre for Environment and Health, London, UK. [99]Cardiovascular Epidemiology and Genetics, Institut Municipal D'investigacio Medica and Centro de Investigación Biomédica en Red CIBER Epidemiología y Salud Pública, Barcelona, Spain. [100]Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland. [101]National Institute for Health and Welfare, Helsinki, Finland. [102]Helsinki University Central Hospital, Unit of General Practice, Helsinki, Finland. [103]Folkhalsan Research Centre, Helsinki, Finland. [104]Vasa Central Hospital, Vasa, Finland. [105]Institute of Genetic Medicine, European Academy Bozen-Bolzano (EURAC), Bolzano-Bozen, Italy, Affiliated Institute of the University of Lübeck, Lübeck, Germany. [106]Department of Neurology, General Central Hospital, Bolzano, Italy. [107]Department of Internal Medicine B, Ernst-Moritz-Arndt University, Greifswald, Germany. [108]Pediatric Endocrinology, Diabetes and Obesity Unit, Department of Pediatrics and Adolescent Medicine, Ulm, Germany. [109]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA. [110]Institut für Klinische Chemie und Laboratoriumsmedizin, Universität Greifswald, Greifswald, Germany. [111]Center for Neurobehavioral Genetics, University of California, Los Angeles, California, USA. [112]Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA. [113]Department of Cardiovascular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, UK. [114]Montreal Heart Institute, Montreal, Quebec, Canada. [115]Department of Medicine III, Pathobiochemistry, University of Dresden, Dresden, Germany. [116]Merck Research Laboratories, Merck and Co., Inc., Boston, Massachusetts, USA. [117]Department of Clinical Sciences, Obstetrics and Gynecology, University of Oulu, Oulu, Finland. [118]National Institute for Health and Welfare, Department of Chronic Disease Prevention, Chronic Disease Epidemiology and Prevention Unit, Helsinki, Finland. [119]MRC Human Genetics Unit, Institute for Genetics and Molecular Medicine, Western General Hospital, Edinburgh, Scotland, UK. [120]Department of Psychiatry and Midwest Alcoholism Research Center, Washington University School of Medicine, St. Louis, Missouri, USA. [121]Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, Regensburg, Germany. [122]Regensburg University Medical Center, Innere Medizin II, Regensburg, Germany. [123]Department of Child and Adolescent Psychiatry, University of Duisburg-Essen, Essen, Germany. [124]Interfaculty Institute for Genetics and Functional Genomics, Ernst-Moritz-Arndt-University Greifswald, Greifswald, Germany. [125]PathWest Laboratory of Western Australia, Department of Molecular Genetics, J Block, QEII Medical Centre, Nedlands, Western Australia, Australia. [126]Busselton Population Medical Research Foundation Inc., Sir Charles Gairdner Hospital, Nedlands, Western Australia, Australia. [127]Division of Research, Kaiser Permanente Northern California, Oakland, California, USA. [128]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA. [129]Department of Social Services and Health Care, Jakobstad, Finland. [130]Core Genotyping Facility, SAIC-Frederick, Inc., National Cancer Institute (NCI)-Frederick, Frederick, Maryland, USA. [131]Institute of Medical Biometry and Epidemiology, University of Marburg, Marburg, Germany. [132]Institut für Epidemiologie und Sozialmedizin, Universität Greifswald, Greifswald, Germany. [133]Research Centre for Prevention and Health, Glostrup University Hospital, Glostrup, Denmark. [134]Faculty of Health Science, University of Copenhagen, Copenhagen, Denmark. [135]National Institute for Health and Welfare, Department of Chronic Disease Prevention, Population Studies Unit, Turku, Finland. [136]Institute of Health Sciences, University of Oulu, Oulu, Finland. [137]Biocenter Oulu, University of Oulu, Oulu, Finland. [138]Hospital for Children and Adolescents, Helsinki University Central Hospital and University of Helsinki, Hospital District of Helsinki and Uusimaa (HUS), Helsinki, Finland. [139]Massachusetts General Hospital (MGH) Weight Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [140]Cardiovascular Research Center and Cardiology Division, Massachusetts General Hospital, Boston, Massachusetts, USA. [141]Framingham Heart Study of the National, Heart, Lung, and Blood Institute and Boston University, Framingham, Massachusetts, USA. [142]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [143]National Institute for Health and Welfare, Diabetes Prevention Unit, Helsinki, Finland. [144]Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. [145]Andrija Stampar School of Public Health, Medical School, University of Zagreb, Zagreb, Croatia. [146]Interdisciplinary Centre for Clinical Research, University of Leipzig, Leipzig, Germany. [147]Department of Medicine, University of Kuopio and Kuopio University Hospital, Kuopio, Finland. [148]Nord-Trøndelag Health Study (HUNT) Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [149]Finnish Institute of Occupational Health, Oulu, Finland. [150]Institut inter-regional pour la santé (IRSA), La Riche, France. [151]Laboratory of Epidemiology, Demography, Biometry, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA. [152]Department of Clinical Chemistry, University of Tampere and Tampere University Hospital, Tampere, Finland. [153]Department of Medicine, Université de Montréal, Montreal, Quebec, Canada. [154]Human Genetics, Genome Institute of Singapore, Singapore, Singapore. [155]Transplantation Laboratory, Haartman Institute, University of Helsinki, Helsinki, Finland. [156]Department of Internal Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [157]Department of Public Health and Primary Care, Institute of Public Health, University of Cambridge, Cambridge, UK. [158]On behalf of the MAGIC (Meta-Analyses of Glucose and Insulin-Related Traits Consortium) investigators. [159]Department of Endocrinology, Diabetology and Nutrition, Bichat-Claude Bernard University Hospital, Assistance Publique des Hôpitaux de Paris, Paris, France. [160]Cardiovascular Genetics Research Unit, Université Henri Poincaré-Nancy 1, Nancy, France. [161]Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Queensland, Australia. [162]Avon Longitudinal Study of Parents and Children (ALSPAC) Laboratory, Department of Social Medicine, University of Bristol, Bristol, UK. [163]Division of Health, Research Board, An Bord Taighde Sláinte, Dublin, Ireland. [164]Institute of Human Genetics, Klinikum rechts der Isar der Technischen Universität München, Munich, Germany. [165]Institute of Human Genetics, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany. [166]Department of Clinical Sciences, Lund University, Malmö, Sweden. [167]Molecular Epidemiology Laboratory, Queensland Institute of Medical Research, Queensland, Australia. [168]Croatian Centre for Global Health, School of Medicine, University of Split, Split, Croatia. [169]Neurogenetics Laboratory, Queensland Institute of Medical Research, Queensland, Australia. [170]National Heart, Lung, and Blood Institute, National Institutes of Health, Framingham, Massachusetts, USA. [171]University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. [172]Department of Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands. [173]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada. [174]Amgen, Cambridge, Massachusetts, USA. [175]Finnish Twin Cohort Study, Department of Public Health, University of Helsinki, Helsinki, Finland. [176]Obesity Research Unit, Department of Psychiatry, Helsinki University Central Hospital, Helsinki, Finland. [177]Department of Medicine, Levanger Hospital, The Nord-Trøndelag Health Trust, Levanger, Norway. [178]Gen-Info Ltd, Zagreb, Croatia. [179]National Institute for Health and Welfare, Oulu, Finland. [180]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK. [181]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland. [182]The Department of Clinical Physiology, Turku University Hospital, Turku, Finland. [183]Clinical Psychology and Psychotherapy, University of Marburg, Marburg, Germany. [184]Department of Clinical Sciences and Clinical Chemistry, University of Oulu, Oulu, Finland. [185]Ludwig-Maximilians-Universität, Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Munich, Germany. [186]South Karelia Central Hospital, Lappeenranta, Finland. [187]Department of Clinical Sciences and Internal Medicine, University of Oulu, Oulu, Finland. [188]Institut für Community Medicine, Greifswald, Germany. [189]Christian-Albrechts-University, University Hospital Schleswig-Holstein, Institute for Clinical Molecular Biology and Department of Internal Medicine I, Kiel, Germany. [190]Universität zu Lübeck, Medizinische Klinik II, Lübeck, Germany. [191]Division of Cardiology, Cardiovascular Laboratory, Helsinki University Central Hospital, Helsinki, Finland. [192]Departments of Medicine and Epidemiology, University of Washington, Seattle, Washington, USA. [193]Department of Psychiatry, Instituut voor Extramuraal Geneeskundig Onderzoek (EMGO) Institute, VU University Medical Center, Amsterdam, The Netherlands. [194]Department of Haematology, University of Cambridge, Cambridge, UK. [195]National Health Service (NHS) Blood and Transplant, Cambridge Centre, Cambridge, UK. [196]Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, UK. [197]Department of Health Sciences, University of Leicester, University Road, Leicester, UK. [198]Department of Medicine, University of Leipzig, Leipzig, Germany. [199]Coordination Centre for Clinical Trials, University of Leipzig, Leipzig, Germany. [200]Department of Medicine, Helsinki University Central Hospital, Helsinki, Finland. [201]Research Program of Molecular Medicine, University of Helsinki, Helsinki, Finland. [202]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [203]Center of Medical Systems Biology, Leiden University Medical Center, Leiden, The Netherlands. [204]Department of Medicine, University of Turku and Turku University Hospital, Turku, Finland. [205]INSERM

Cardiovascular Genetics team, Centre Investigation Clinique (CIC) 9501, Nancy, France. [206]Steno Diabetes Center, Gentofte, Denmark. [207]Center for Human Genomics, Wake Forest University, Winston-Salem, North Carolina, USA. [208]Department of Physiatrics, Lapland Central Hospital, Rovaniemi, Finland. [209]School of Pathology and Laboratory Medicine, University of Western Australia, Nedlands, Western Australia, Australia. [210]Department of Internal Medicine, University of Oulu, Oulu, Finland. [211]School of Medicine and Pharmacology, University of Western Australia, Perth, Western Australia, Australia. [212]Department of Clinical Physiology, University of Tampere and Tampere University Hospital, Tampere, Finland. [213]Stanford University School of Medicine, Stanford, California, USA. [214]Department of Psychiatry and Human Behavior, University of California, Irvine (UCI), Irvine, California, USA. [215]Leipziger Interdisziplinärer Forschungs-komplex zu molekularen Ursachen umwelt- und lebensstilassoziierter Erkrankungen (LIFE) Study Centre, University of Leipzig, Leipzig, Germany. [216]Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois (CHUV) University Hospital, Lausanne, Switzerland. [217]Human Genetics Center and Institute of Molecular Medicine, University of Texas Health Science Center, Houston, Texas, USA. [218]Faculty of Health Science, University of Southern Denmark, Odense, Denmark. [219]New York University Medical Center, New York, New York, USA. [220]National Institute for Health and Welfare, Department of Mental Health and Substance Abuse Services, Unit for Child and Adolescent Mental Health, Helsinki, Finland. [221]NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, UK. [222]Department of Urology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [223]Institute of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark. [224]Faculty of Health Science, University of Aarhus, Aarhus, Denmark. [225]Department of Psychiatry, Leiden University Medical Centre, Leiden, The Netherlands. [226]Department of Psychiatry, University Medical Centre Groningen, Groningen, The Netherlands. [227]Department of Neurology, University of Lübeck, Lübeck, Germany. [228]Institute for Paediatric Nutrition Medicine, Vestische Hospital for Children and Adolescents, University of Witten-Herdecke, Datteln, Germany. [229]Department of Medicine III, Prevention and Care of Diabetes, University of Dresden, Dresden, Germany. [230]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, Maryland, USA. [231]Hjelt Institute, Department of Public Health, University of Helsinki, Helsinki, Finland. [232]South Ostrobothnia Central Hospital, Seinajoki, Finland. [233]Department of Internal Medicine, Centre Hospitalier Universitaire Vaudois (CHUV) University Hospital, Lausanne, Switzerland. [234]The Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. [235]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, USA. [236]Pacific Biosciences, Menlo Park, California, USA. [237]Sage Bionetworks, Seattle, Washington, USA. [238]Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri, USA. [239]Division of Intramural Research, National Heart, Lung, and Blood Institute, Framingham Heart Study, Framingham, Massachusetts, USA. [240]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, New York, USA. [241]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [242]Department of Medical Genetics, University of Helsinki, Helsinki, Finland. [243]Laboratory of Genetics, National Institute on Aging, Baltimore, Maryland, USA. [244]Division of Community Health Sciences, St. George's, University of London, London, UK. [245]Klinikum Grosshadern, Munich, Germany. [246]Faculty of Medicine, University of Iceland, Reykjavík, Iceland. [247]University of Cambridge Metabolic Research Labs, Institute of Metabolic Science Addenbrooke's Hospital, Cambridge, UK. [248]Carolina Center for Genome Sciences, School of Public Health, University of North Carolina Chapel Hill, Chapel Hill, North Carolina, USA. [249]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [250]These authors contributed equally to this work. Correspondence should be addressed to M.B. (boehnke@umich.edu), K. Stefansson (kstefans@decode.is), K.E.N. (kari_north@unc.edu), M.I.M. (mark.mccarthy@drl.ox.ac.uk), J.N.H. (joelh@broadinstitute.org), E.I. (erik.ingelsson@ki.se) or R.J.F.L. (ruth.loos@mrc-epid.cam.ac.uk).

ONLINE METHODS

Study design. We designed a multistage study (**Supplementary Fig. 1**) comprising a genome-wide association meta-analysis (stage 1) of data on up to 123,865 genotyped individuals from 46 studies and selected 42 SNPs with $P < 5 \times 10^{-6}$ for follow up in stage 2. Stage 2 comprised up to 125,931 additional genotyped individuals from 42 studies. Meta-analysis of stage 1 and stage 2 summary statistics identified 32 SNPs that reached genome-wide significance ($P < 5 \times 10^{-8}$).

Stage 1 genome-wide association meta-analysis. *Samples and genotyping.* The GIANT consortium currently encompasses 46 studies with up to 123,865 genotyped adult individuals of European ancestry with data on BMI (**Supplementary Note**). The samples from 46 studies, including between 276 and 26,799 individuals each, were genotyped using Affymetrix and Illumina whole genome genotyping arrays (**Supplementary Note**). To allow for meta-analysis across different marker sets, imputation of polymorphic HapMap European CEU SNPs (**Supplementary Note**) was performed using MACH[45], IMPUTE[46] or BimBam[47].

Association analysis with BMI. Each study performed single marker association analyses with BMI using an additive genetic model implemented in MACH2QTL (Y. Li, C.J.W., P.S. Ding and G.R.A., unpublished data), Merlin[48], SNPTEST[46], ProbAbel[49], GenABEL[50], LME in R or PLINK[51]. BMI was adjusted for age, age^2 and other appropriate covariates (for example, principal components) and inverse normally transformed to a mean of 0 and a standard deviation of 1. Analyses were stratified by sex and case status (for samples ascertained for other diseases) (**Supplementary Note**). To allow for relatedness in the SardiNIA, Framingham Heart, Amish HAPI Heart and Family Heart studies, regression coefficients were estimated in the context of a variance component model that modeled relatedness in men and women combined with sex as a covariate. Before meta-analyzing the genome-wide association data for the 46 studies, SNPs with poor imputation quality scores (r^2.hat < 0.3 in MACH, observed/expected dosage variance < 0.3 in BimBam or proper_info < 0.4 in IMPUTE) and those with a minor allele count (MAC = $2 \times N \times$ minor allele frequency) of < 6 in each sex- and case-specific stratum were excluded for each study. All individual GWAS were genomic control corrected before meta-analysis. Individual study-specific genomic control values ranged from 0.983 to 1.104 (**Supplementary Note**).

Meta-analysis of stage 1 association results. Next, we performed the stage 1 meta-analysis using the inverse variance method, which is based on β values and standard errors from each individual GWAS. To ensure consistency of results, we also performed the stage 1 meta-analysis using the weighted *z*-score method, which is based on the direction of association and *P* values of each of the individual studies. Both meta-analyses were performed using METAL (see URLs), and the correlation between the resulting $-\log_{10} P$ values was high ($r >$ 0.99). For the discovery of replicating variants, the results of the inverse variance meta-analysis were used followed by a final genomic control correction of the meta-analyzed results. The genomic control value for the meta-analyzed results before genomic control correction was 1.318.

Selection of SNPs for follow up. Forty-two lead SNPs, representing the forty-two most significant ($P < 5 \times 10^{-6}$) independent loci, were selected for replication analyses (stage 2) (**Supplementary Table 1**). Loci were considered independent when separated by at least 1 Mb. For some loci, the SNP with the strongest association could not be genotyped for technical reasons and was substituted by a proxy SNP that was in high LD with it ($r^2 > 0.8$) according to the HapMap CEU data (**Supplementary Table 1**). We tested the association of these 42 SNPs in 16 *de novo* and 18 *in silico* replication studies in stage 2.

Stage 2 follow up. *Samples and genotyping.* Directly genotyped data for the 42 SNPs was available from a total of 79,561 adults of European ancestry from 16 studies using Sequenom iPLEX or TaqMan assays (**Supplementary Note**). Samples and SNPs that did not meet the quality control criteria defined by each individual study were excluded. Minimum genotyping quality control criteria were defined as Hardy-Weinberg Equilibrium $P > 10^{-6}$, call rate > 90% and concordance > 99% in duplicate samples in each of the follow-up studies. Association results were also obtained for the 42 most significant SNPs from 46,370 individuals of European ancestry from 18 GWAS that had not been included in the stage 1 analyses (**Supplementary Note**). Studies included between 345 and 22,888 individuals and were genotyped using Affymetrix and Illumina genome-wide genotyping arrays. Autosomal HapMap SNPs were imputed using either MACH[45] or IMPUTE[46]. SNPs with poor imputation quality scores from the *in silico* studies (r^2.hat < 0.3 in MACH or proper_info < 0.4 in IMPUTE), and SNPs with a MAC < 6 in each sex- and case-specific stratum were excluded.

Association analyses and meta-analysis. We tested the association between the 42 SNPs and BMI in each *in silico* and *de novo* stage 2 study separately as described for the stage 1 studies. We subsequently meta-analyzed β values and standard errors from the stage 2 studies using the inverse-variance method. The meta-analysis using a weighted *z*-score method was similar (the *r* between *P* values was >0.99) and included up to 249,796 individuals. Data was available for at least 179,000 individuals for 41 of the 42 SNPs. For one SNP (rs6955651 in *KIAA1505*), data was only available for 125,672 individuals due to technical challenges relating to the genotyping and imputation of this SNP. Next, we meta-analyzed the summary statistics of the stage 1 and stage 2 meta-analyses using the inverse-variance method in METAL.

Assessment of population stratification. To assess for possible inflation of test statistics by population stratification, we performed a family-based analysis, which is immune to stratification, in 5,507 individuals with pedigree information from the Framingham Heart Study using that the QFAM–within procedure in PLINK. Effect sizes and directions in the Framingham Heart Study data are the β statistics reported by PLINK from the within-family analysis, and the *P* values are empirical and are based on permutation testing. For imputed SNPs, only those with r^2.hat > 0.3 in MACH were analyzed using the best-guess genotypes from dosages reported by MACH. For the 32 loci in general and the 18 new loci in particular, the estimated effect sizes on BMI were essentially identical in the overall meta-analysis and in the Framingham Heart Study sample (**Supplementary Note**), and, as expected in the absence of substantial stratification, about half of the loci (18 out of 32 loci total and 10 out of 18 new loci) had a larger effect size in the family-based sample. These results indicate that the genome-wide significant associations in our meta-analysis are not substantially confounded by stratification.

In addition, we estimated the fixation index (Fst) for all SNPs to test whether the 32 confirmed BMI SNPs might be false-positive results due to population stratification. We selected five diverse European populations with relatively large sample sizes (Northern Finland Birth Cohort (NFBC), British 1958 Birth Cohort, SardiNIA, CoLaus and DeCODE) for this analysis. The mean Fst value for the 32 confirmed BMI SNPs was not significantly different from the mean Fst for 2.1 million non-BMI associated SNPs (*t* test $P = 0.28$), suggesting that the SNPs that are associated with BMI do not appear to have strong allele frequency differences across the European samples examined.

Follow-up analyses. Subsequently, we performed an extensive series of follow-up analyses to estimate the impact of the 32 confirmed BMI loci in adults and children and to explore their potential functional roles. These follow-up analyses are described in detail in the **Supplementary Note**.

In brief, we estimated the cumulative effect of the 32 loci combined on BMI and assessed their predictive ability in obesity and BMI in the ARIC study. Association between the 32 confirmed BMI variants and overweight or obese status was assessed in stage 2 samples, and association with BMI in children and adolescents was examined in four population-based studies. Furthermore, we tested for association between the 32 SNPs and extreme or early-onset obesity in seven case-control studies of extremely obese adults and extremely obese children or adolescents. Data on the association between the 32 SNPs and height and weight were obtained from the stage 2 replication samples, and data on the association with related traits were extracted from previously reported genome-wide association meta-analyses for type 2 diabetes (Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium[18]), lipid levels (the Global Lipids

Genetics Consortium[20]) and glycemic traits (Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)[19,21]).

To discover potentially new pathways associated with BMI, we tested whether predefined biological processes or molecular functions that contain at least one gene within 300 kb of the 32 confirmed BMI SNPs were enriched for multiple modest BMI associations using MAGENTA[33]. We identified SNPs having $r^2 \geq 0.75$ with the lead SNP that were likely non-synonymous, nonsense or which occurred within 5 bp of the exon-intron boundary and also evaluated whether any of the 32 confirmed BMI SNPs tagged common CNVs. We examined the *cis* associations between each of the 32 confirmed BMI SNPs and expression of nearby genes in adipose tissue[34,52], whole blood[34], lymphocytes[36,52] and brain[35].

We evaluated the amount of phenotypic variance explained by the 32 BMI loci using a method proposed by the International Schizophrenia Consortium[37] and estimated the number of susceptibility loci that are likely to exist using a new method[38] based on the distribution of effect sizes and minor allele frequencies observed for the established BMI loci and the power to detect those effects in the combined stage 1 and stage 2 analysis.

We performed a conditional genome-wide association analysis to examine whether any of the 32 confirmed BMI loci harbored additional independent signals, and we also examined gene-by-gene and gene-by-sex interactions among the BMI loci. Dominant and recessive analyses were performed for the 32 confirmed BMI SNPs to test for non-additive effects.

45. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
46. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
47. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet.* **4**, e1000279 (2008).
48. Abecasis, G.R. & Wigginton, J.E. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet.* **77**, 754–767 (2005).
49. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134 (2010).
50. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
51. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
52. Zhong, H., Yang, X., Kaplan, L.M., Molony, C. & Schadt, E.E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* **86**, 581–591 (2010).