M.V. STRUCHALIN

APPROACHES TO DISSECT THE COMPLEX GENETIC ARCHITECTURE OF COMMON TRAITS

Financial support by the Centre for Medical Systems Biology (CMSB) and Erasmus University for publication of this thesis is gratefully acknowledged.

Struchalin, Maksim V.
Approaches to dissect the complex genetic architecture of common traits
M.V. Struchalin - Rotterdam
Erasmus University Medical Center
Deptartment of Epidemiology
ISBN: 978-XX-XXX-XXXX-X

APPROACHES TO DISSECT THE COMPLEX GENETIC ARCHITECTURE OF COMMON TRAITS

Benaderingen voor onderzoek naar de complexe genetische achtergrond
van veelvoorkomende eigenschappen

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de erasmus universiteit rotterdam
op gezag van de rector magnificus
prof.dr. h.g. schmidt
en volgens besluit van het college voor promoties.

de openbare verdediging zal plaatsvinden op
woensdag 22 mei 2013 om 9:30 uur

door

m.v. struchalin

geboren te Novosibirsk, Russia

ERASMUS UNIVERSITEIT ROTTERDAM

Promotoren:	Prof.dr.ir. C.M. van Duijn
	Prof.dr. B.A. Oostra
Overige leden:	Prof.dr. O.H. Franco Duran
	Prof.dr. M.H. Kayser
	Prof.dr. A.G. Uitterlinden
Copromotor:	Dr. L.C. Karssen

PUBLICATIONS AND MANUSCRIPTS BASED ON THE STUDIES DESCRIBED IN THIS THESIS

Chapter 2.1

Aulchenko YS, **Struchalin M**, van Duijn CM, *ProbABEL package for genome-wide association analysis of imputed data*, BMC Bioinformatics. 2010 Mar 16;11:134.

Chapter 2.2

Liu F, **Struchalin M**, Duijn K, Hofman A, Uitterlinden AG, Duijn C, Aulchenko YS, Kayser M, *Detecting low frequent loss-of-function alleles in genome wide association studies with red hair color as example*, PLoS One. 2011;6(11)

Chapter 2.3

Struchalin M, Dehghan A, Witteman JC, van Duijn C, Aulchenko YS, *Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations*, BMC Genet. 2010 Oct 13;11:92.

Chapter 2.4

Struchalin M, Amin N, Eilers PH, van Duijn CM, Aulchenko YS, *An R package "VariABEL" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity*, BMC Genet. 2012 Jan 24;13:4.

Chapter 3.1

Aulchenko YS*, **Struchalin M***, Belonogova NM, Axenovich TI, Weedon MN, Hofman A, Uitterlinden AG, Kayser M, Oostra BA, van Duijn CM, Janssens AC, Borodin PM, *Predicting human height by Victorian and genomic methods*, Eur J Hum Genet. 2009 Aug;17(8):1070-5

Chapter 3.2

Struchalin M, Lennart C Karssen, Najaf Amin, Kelly S Benke, Abbas Dehghan, Jacqueline C Witteman, Albert Hofman, Ben A Oostra, Oscar H Franco Duran, Cornelia M van Duijn, *The role of common genetic and environmental factors in extreme high and low levels of total cholesterol*. (In preparation)

Chapter 3.3

E.G. van den Herik*, **Struchalin M***, L.M.L. de Lau, H.M. den Hertog, S. Fonville, P.J. Koudstaal, C.M. van Duijn, *Associations between recently discovered genetic variations in metabolic traits and arterial stenosis in patients with recent cerebral ischemia.* (In preparation)

* *Those authors contributed equally*

CONTENTS

1

CHAPTER: GENERAL INTRODUCTION

The fundamental aim of analysis in statistical genetics is to establish the link between phenotypes and genotypes. The first successful method that allowed identification of genomic regions influencing a trait of interest was linkage analysis [1]. The idea behind the method is to test how often a trait co-segregates with a specific genomic region. This method gained wide popularity due to the advent of two new techniques for genotyping polymorphic loci. One of them is restriction fragment length polymorphism (RFLP) [2] which became possible in 1970s due to the discovery of restriction enzymes that cut DNA at specific nucleotide sequences [3]. Another technique, developed later in 1985, is polymerase chain reaction (PCR) which allows *in vitro* amplification of short segments of DNA from a template using DNA polymerase [4]. Later, PCR was used to perform the genotyping of microsatellites – repeating sequences of nucleotides which have a high level of polymorphism in the human genome [5]. RFLPs and microsatellites were widely used in linkage analysis to localize loci responsible for Mendelian traits (traits that are controlled by a single locus or a small number of loci and having high penetrance). For the next 15 years linkage analysis was an important method in genetic epidemiology facilitating the discovery of more than a thousand new loci [2].

Linkage analysis is the method of choice for identification of regions harboring rare, high risk mutations [2], basically making it applicable to traits controlled by a single locus or a small number of loci containing alleles with a large effect size. Indeed, Mendelian traits, while rare in the general population, are frequent in families where the rare mutation segregates with high penetrance. This allows for a very effective design by sampling families where the disease segregates via the proband. Close relatives in such a family will share a very large proportion of their genome identical-by-descent, and therefore a few hundreds of genetic markers allow tagging the co-segregation of the disease with a region of the genome.

By the middle of the 1990s it became evident that linkage, while being successful for analysis of many monogenic diseases, is less effective for complex traits, which are controlled by many loci each having modest effect. By the end of the 1990s, two hypotheses about the genetic control of complex traits gained popularity. One of them states that most complex traits (including quantitative traits and common diseases) are controlled by common alleles of small effect (*common disease / common variant* hypothesis). The other hypothesis states that common traits are controlled by many rare alleles with relatively high effects (*common disease/multiple rare variants* hypothesis). Risch and Merikangas showed in their work [6] that under a *common disease / common variant* model the association analysis is the more powerful method in comparison with linkage analysis. Such analysis requires testing of the correlation between a genetic variant and the trait or the disease of interest. It has been proposed that genome-wide association analysis in case-control or population-based samples can serve the needs of identification of loci controlling complex traits. However, unrelated individuals are sharing only small parts of their genome identical-by-descent. Consequently, successful Genome-Wide Association Studies (GWAS) require the development of very dense marker maps.

By the beginning of the year 2000 many laboratories contributed to the development of the Single Nucleotide Polymorphisms (SNPs) map. SNPs are genetic markers which are currently used in GWAS for identification of the genetic variation responsible for the variation of a phenotype of interest. A SNP is a nucleotide in the DNA sequence which differs between members in a population. Most SNPs present variation of two nucleotides, however, there are SNPs known with three or more nucleotides. Figure 1.1 illustrates an example of a SNP with two variants (A and G) in a region of one homologous chromosome of seven individuals. For the statistical power in genome-wide association analysis it is important that SNPs are common and can facilitate reconstructing the haplotype where it resides in. A haplotype is a set of SNPs on a chromosome that are correlated to each other. In 2001, The SNP Consortium and The International Human Genome Sequencing Consortium described a map of 1.42 million SNPs [7]. The paper of C. Venter (who was a founder of the private company Celera Genomics which performed the sequencing of the human genome independently from the public effort led by by Francis Collins, The Human Genome Project) that was published on the same day, describes a map of 2.1 million SNPs [8]. To summarize and to extend the SNP data, the HapMap project was started in 2002. The HapMap is an international collaboration among many researchers which additional to the SNP was aimed to develop a haplotype map. The basic idea of HapMap was to discover how SNPs are organized on chromosomes and how these combinations (haplotypes) segregate together in various populations. All data from the HapMap project was placed in the public domain and made available for download. The first version of HapMap which was released in 2005 contained approximately 1 million SNPs [9] and was based on the genetic information from 269 individuals from four geographically diverse populations (African, European, Japanese and Chinese). The second-generation map increased this number to 3.1 million SNPs in 2007 [10]. In 2010, in the last version of HapMap a map of 1.6 million SNPs was reported for an increased number of individuals (1, 184 individuals from 11 populations) [11]. However, many genetic variants, in particular the rare ones, have remained undiscovered in HapMap. In 2012, the 1000 Genomes Project which was launched in 2008 to build an even more detailed map of human genome variation (including in particular more rare variants) released a haplotype map of 38 million SNPs based on 1, 092 individuals from 14 populations [12].

Mapping of SNPs in the human genome is very important for studying genetic associations with traits and diseases. Knowledge of the map of SNPs in the human genome facilitated the development of DNA arrays for massive parallel

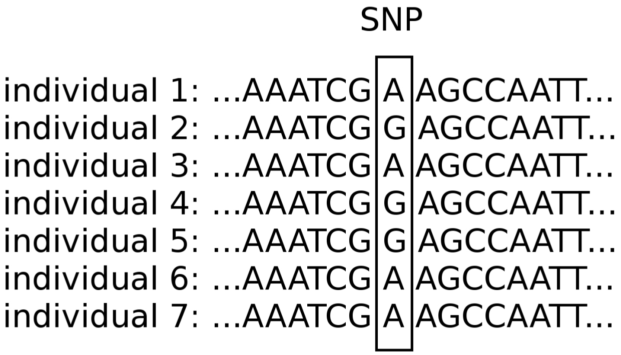

SNP
individual 1: ...AAATCG A AGCCAATT...
individual 2: ...AAATCG G AGCCAATT...
individual 3: ...AAATCG A AGCCAATT...
individual 4: ...AAATCG G AGCCAATT...
individual 5: ...AAATCG G AGCCAATT...
individual 6: ...AAATCG A AGCCAATT...
individual 7: ...AAATCG A AGCCAATT...

Figure 1.1: Illustration of a SNP having A and G variants. A region of one homologous chromosome is presented.

genotyping. The first array for genotyping of more than 100,000 SNPs was released in 2004 [13]. By 2005, genotyping of hundreds of thousands of SNPs became affordable for many research groups world-wide.

Additionally, HapMap is used to reduce the number of SNPs that need to be genotyped by allowing imputation of untyped SNPs, thus decreasing the cost of each single GWAS. The methods which are used for imputation of untyped SNPs are based on searching common haplotypes between an individual's genome and a reference panel with a high density of genotyped SNPs such as HapMap and inferring missing genotypes from common haplotypes found in the reference. These methods can reliably impute up to 10 million SNPs using only 500,000 typed SNPs and haplotype information from the latest haplotype map from the 1000 Genomes Project.

Nowadays, most GWAS findings are obtained in the framework of big consortia efforts where the use of HapMap information (as a reference panel for SNP imputation) plays a crucial role. GWAS results obtained in different population studies are meta-analyzed together in a consortium, thus increasing the power to detect smaller SNP effects. However, different studies rely on different genotyping platforms which may have little overlap in SNP content. For example, the Illumina 317K array and the Affymetrix 500K array have only approximately 51,000 SNPs in common. The imputation procedure can provide a common panel of SNPs for each study making meta-analysis possible.

The first successful GWAS was done in 2004 using a newly-developed genotyping array [13]. It revealed one locus — Complement factor H (*CFH*) — associated with age-related macular degeneration [14]. By 7 December 2012, according to the "GWAS Integrator" [15], 1381 GWASs of 738 phenotypes have been published, reporting about 7192 SNP associations. Without a doubt GWAS has become an important tool for studying the genetic architecture of complex traits.

Imputation of untyped SNPs plays an important role in genome-wide associations studies and a number of methods have been developed [16, 17, 18]. Implementation of imputation methods results in estimates of the posterior probability distributions $P_g = (P_{AA}, P_{AB}, P_{BB})$ of the genotypes based on the available data. For many genomic loci, this distribution may be non-degenerate.

Several techniques can be applied to the analysis of such "uncertain" data. The simplest approach would be to use the "best-guess genotypes", i.e., to use the genotype with the highest posterior probability ($g = \max_g P_g$) for analysis as if it were a directly-typed marker. This approach is equivalent to replacing the estimated probability distribution with a degenerate one where a probability of 1 is assigned to the genotype with the maximal posterior probability. From standard statistical theory it is known, however, that such a procedure results in biased estimates of the effects and, consequently, to loss of power. A correct analysis can be achieved using a maximum likelihood approach. Under this approach the likelihood can be computed using the total probability formula in which summation is performed over the genotypes, whose true values are not known, but whose posterior probabilities can be estimated given the data. This approach is computationally demanding, as it requires summation over the underlying probability distribution and numerical maximization of the likelihood function. Alternatively, a regression approach in which the posterior genotypic probabilities are used as independent variables can be applied. The main advantage of this approach is that well-established regression analysis methodology, algorithms, and code can be used in its implementation. In Chapter 2.1, the `ProbABEL` package is described, which is designed to perform genome-wide regression on posterior genotypic probabilities in a computationally efficient manner.

One of the very important aspects to consider in any GWAS is confounding. Confounding factors are associated with both the trait and the risk factor under investigation. In GWAS, where association between traits and genetic markers is studied, the genetic origin of study participants may work as a major confounder. The most common confounding factor is population stratification which can occur, for example, when a population consists of a number of ethnically different subpopulations. Consider for example the case where a GWAS of height is conducted in a sample where two different populations (e.g., Dutch and Chinese) are analyzed jointly. Due to genetic drift and different ancestry these populations have substantial genetic differences that are particularly manifested in different allelic frequencies for many SNPs. Additionally, those populations have a different average height value (Dutch people are taller than Chinese people). If we analyze an association between a SNP which, for example, is represented mostly by allele A in Dutch people and by allele C in Chinese people, we will see that tall people have mostly allele A and short people have mostly allele C. Consequently, the erroneous conclusion about presence of association between this SNP and height can be made if the confounding factor (population stratification) is not taken into consideration.

The same effect, albeit less pronounced, is observed for the samples containing genetically closely related individuals. This is the case when a sample is collected from genetically isolated populations or families. The confounding factor is the common ancestry that determines genetic similarity of subsamples and similarity of a studied trait which is due to genetic and environmental factors.

Different methods can be applied to control for confounding caused by genetic factors. In the case of population stratification the approach widely used today is to perform the GWAS in each population separately and then combine the obtained results in a meta-analysis. In the case of high ethnic heterogeneity of the sample the principle component analysis is used to adjust for the confounding due to population structure. When analyzing a sample containing families, mixed models and the two-step score test approximation to the mixed model as described in Chapter 2.1 are currently used.

NOVEL METHODS AND SOFTWARE TO DISSECT HERITABILITY OF COMPLEX TRAITS

Detecting low frequent loss-of-function alleles in genome-wide association studies

GWAS methodology was developed to identify effects of common variants which were expected to be responsible on a substantial proportion of heritability. However, the recent achievement in GWAS demonstrated that those expectation are not fully met. For the most of common traits, the common variants identified in GWAS explain a relatively small proportion of heritability. This phenomenon is usually referred to as the "hidden heritability" or "missing heritability" [19].

Most GWASs performed up until now assumed an additive model of association between the phenotypes and the investigated genotypes, however, there are reasons to believe that recessive models of control may be wide-spread at least for some traits. A recent study on height [20] on genome-wide recessive effects showed highly significant association between height and genome-wide homozygosity. Moreover, evolutionary reasoning predicts that recessive mutations can reach higher frequency, and have larger effect compared to dominant or additive mutations. It is known that some recessive disorders can be determined by two unrelated recessive alleles located in the same locus which, however, are in heterozygous state. This condition is called 'compound heterozygosity' and the known example of this is the cystic fibrosis [21].

It has been proposed that heterozygous loss-of-function (LOF) variants may account for the essential proportion of heritability [19, 22]. LOF variants represent alleles resulting in reduced or loss of protein function by disrupting not only the protein-coding genes but also any essential genetic element, including non-coding regulatory motifs. They have a variety of forms, including single-base substitutions such as nonsense SNPs, splice site disruptions and small or larger

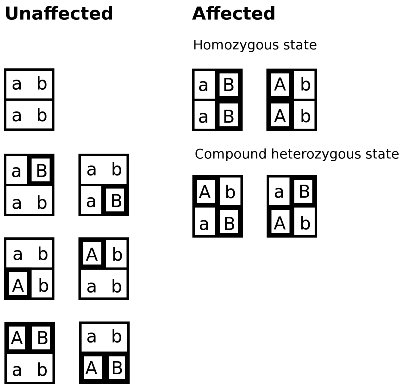


Figure 1.2: Illustration of how recessive LOF variants in CH state influence phenotype. Each square on the figure shows an individual's haplotype. Affected individuals have two risk variants (capital A or B in a bold square), one on each homologous chromosome.

insertions/deletions that change the reading frame or remove an entire gene. LOF variants are mostly recognized by their genetic association with a variety of phenotypes largely inherited in a recessive manner. It is important to note that multiple LOF variants at the same locus can act not only in the homozygous state, but also in the compound heterozygous (CH) state, where the presence of two different LOF variant alleles at the same gene, one on each homologue chromosome, influences the phenotype (Figure 1.2). In such cases, the CH state would be much more frequent than the homozygous state for any individual variant. This scenario presents a specific manifestation of interaction between variants. Standard GWAS would have limited power to detect loci having the compound heterozygosity architecture. In Chapter 2.2, a novel method of testing for association between LOF in the CH stage and a trait is described.

Detecting interacting genetic variants in genome-wide association studies

Most GWASs are focused on testing association between a single SNP and a phenotype. One of the mechanisms which could explain a larger proportion of the phenotypic variance is interaction between two or more variants or between variants and environmental factors. Several methods have been proposed to search for interacting loci. For example, linear or logistic regression where the interaction term is included into a model. However, at least for a logistic regression, it was shown that the power of such analysis is low [23]. For case-control studies a number of methods have been developed with higher power compared to logistic regression (which are summarized in [23]). These methods use the assumption of independence between SNP and environmental factors.

However, violation of that assumption results into a high rate of false positives. Some of these methods try to protect against false positives but still have inflated type I error.

The methods described above are suitable for testing of interaction between a single gene and environmental factors. In the case of gene-gene interaction where millions of SNPs need to be considered, as has become routine in GWAS nowadays, testing for interaction for all possible combinations of SNPs becomes cumbersome requiring parallel computations using hundreds or thousands of CPU cores - even if only the "simplest" case of pairwise SNP interactions is considered. Moreover, a large number of models have to be tested, resulting in a multiple testing problem, which weakens the statistical power and consequently the possibility of new findings.

In Chapter 2.3 I describe a novel method for discovering potentially interacting loci in genome-wide scans. The idea behind this method is to test the variance difference of a trait between different genotypic groups. In the presence of interaction between a given SNP and another genetic or non-genetic factor the phenotypic variance is expected to differ between genotypic groups of a given SNP. In this case, to detect an interacting SNP, each SNP available in a study is tested only once. The method answers a question which SNP is potentially involved into interaction with some factor (genetic or non-genetic) and it is not necessary to know the factor itself.

However, the initial methods proposed both by me in Chapter 2.3 and by others [24] cannot be used for the analysis of imputed data which is crucial for achieving large sample sizes and hence power by the means of joint/meta-analysis of multiple data-sets. Therefore, next I extended the method for analysis of imputed SNPs. In Chapter 2.4, I describe the extension of the method. The idea of testing genotypic variance difference reduces to testing the mean difference of the squared trait with prior normalization of the mean difference. We implemented this new approach in a software package called `VariABEL`. The software is written in the R language, uses compiled code written in C/C++ and belongs to the `GenABEL` suite. This implies complete compatibility with the previous products of the suite including the widely used packages for GWAS as `GenABEL` and `ProbABEL`.

The method of interaction testing through testing of variance difference is gaining popularity in the scientific community. A recent study on BMI and height [25] where meta-analysis of genome-wide association studies of phenotypic variation using $\approx 170,000$ individuals was performed reported the *FTO* gene in which SNP rs7202116 is associated with phenotypic variability. The variance difference between the two opposite genotypes is 7% showing the great potential of this method.

POST-GWAS STUDIES TO DISSECT COMPLEX GENETIC ARCHITECTURE OF COMMON TRAITS

Prediction of phenotypes with example on height

Despite the fact that our knowledge about genetic basis of common traits is incomplete, a very important development in epidemiology is to translate GWAS findings into prediction of human traits and the onset or progression of diseases. This will be crucial for preventive strategies but moreover, accurate prediction of human traits opens broad opportunities in forensics. Genetic material (obtained from, e.g., blood spots) left by a person on a crime scene can provide valuable information about the height, eye color or facial features of the person. In medicine, such prediction can answer the question whether an individual has, for example, high lipid levels because of genetic or environmental factors. If the individual has low predicted genetic risk of high lipid levels and, at the same time, has high lipid levels then it is more likely that this individual needs to follow specific regimens like a balanced diet or physical exercises instead of taking lipid lowering medicine like statins. Prediction of the genetic risk to develop a disease for a particular person can enable targeted preventative treatments that can eliminate or, at least, reduce manifestation of the disease.

It is worth noting that trait prediction is also valuable in other fields such as animal or plant breeding. These fields aim to improve various features such as meat, milk or growth rate of livestock or increased yields in crops. There are possibly many features of animals and plants which have a complex genetic architecture involving many genetic factors, some of which may interact with each other or with environmental factors. Accurate prediction of such features can substantially facilitate the breeding process.

Any prediction study in epidemiology starts from developing a predictive regression model (a set of predictors and mutual relationship between them). The predictors can be divided in two categories. The first category contains environmental factors such as individual's age, gender, smoking status, blood pressure, lipids levels and so on. Strictly speaking, some of those factors can not be considered as purely environmental (non-genetic) as they them-self have a genetic component (e.g., blood pressure, lipids levels). However, incomplete knowledge about their genetic basis and obvious strong effect of some of those factors on a studied trait make them useful for inclusion in the predictive model. The second category represents genetic factors known to be associated with a trait of interest. Nowadays, a polygenic model is commonly used for prediction of a trait's genetic component. In such a model, the effect of each variant is assumed to be additive (the effect of a SNP is proportional to a number of risk variants) and the prediction of the trait value is simply the summation of the effects of all the genetic variants. This model is called the weighed genetic (allelic) risk score. A similar model called the unweighted genetic (allelic) risk score is used, the only difference being that the variants are supposed to influence the trait with the same effect size. There are studies showing presence of interaction between environmental and genetic factors. This can be reflected in the predictive model by inclusion of the interaction term. The estimation of the predictor's effect (known as calibration of a predictive model) is conducted in a discovery (training) data set. Effect of environmental factors are obtained from the regression analysis and effects of the genetic variants – in GWASs. In the stage of estimating predictor's effect, the caution should be taken during developing the model to avoid inclusion of highly correlated predictors. The phenomenon when such predictors are present in the model is called multicollinearity. This results in incorrect estimation of predictors effect that can decrease predictive power of the model.

The next important step after calibration of a predictive model is validation. The validation is necessary for estimating an accuracy of prediction of future outcomes (trait's values). The model can be validated on the same data set which was used for the development of this model. This is a fast and simple procedure which, however, gives an overestimated predictive accuracy that can result in the wrong conclusions about performance of the predictive model. A reason for overestimation can be an overfitting – fitting a small data set by an excessively complex predictive model with many predictors. In this case the model describes a random error or noise instead of the underlying relationship. The overfitting can make it difficult to compare several predictive models which differ in a number of predictors even if their maximum number is relatively small. In this case, even if none of these predictors are associated with a trait, the model with the highest number of predictors will show the highest predictive power and can be wrongly chosen as the best among other tested. The commonly used methods for validation of predictive model in statistical genetics which lack such disadvantage are cross-validation and bootstrapping. Those methods use different subsamples from the original sample for development and validation of a predictive model. This approach can protect against overfitting but still can give an overestimated predictive accuracy. The reason for this is that the predictive model can be specific for the training data set used for the calibration of the model. An example of such specific predictive model can be a hypothetical model where genetic variants specific for a given population are used as predictors. This model will obviously show a lower performance if being applied on the data from population where such genetic variants does not exist or have no influence on the trait. In this example, using a different population will protect against this effect and give unbiased estimation of the predictive accuracy.

The accuracy of phenotype or disease prediction depends on how complete our knowledge is about its genetic and environmental background. For quantitative traits, the accuracy is expressed as a proportion of the total variance that the predictors explain in a population. This proportion is determined as squared covariation between observed and predicted trait divided by magnitudes of their variances:

$$r^2 = \frac{\text{cov}(\text{trait}, \text{prediction})^2}{(\text{var}(\text{trait}) \times \text{var}(\text{prediction}))}. \tag{1.1}$$

The parameter r^2 shows not only the accuracy of the prediction at the population level but also allows estimating an average error in the trait prediction for each individual in the population. This error is given by

$$(1 - r^2) \times \sigma^2_{\text{trait}}, \tag{1.2}$$

where σ^2_{trait} is the variance of the trait in the population to which the individual belongs.

For binary traits, a different metric is used to characterize prediction accuracy. The outcome of prediction for a binary trait is the probability for an individual to develop the disease. For a practical application of such information (in, for example, medicine) many end-users are more interested in the answer to the question whether the person is considered to have a high risk to develop the disease (and, therefore, should receive treatment) or not. To give the answer based on the probabilistic outcome from a prediction test the user has to set up a threshold T (a value ranged from 0 to 1) for which the individual is considered to have a high risk if the probability to develop the disease exceeds this threshold. Predictive tests often have limited accuracy leading to misclassification of an individual's risk. There are two measures of classification accuracy: sensitivity and specificity. Sensitivity is the proportion of individuals who were classified as having high risk and, actually developed the disease. In statistical theory this is called the proportion of true positives. Specificity is the proportion of individuals who were classified as having low risk to develop the disease and who truly stay healthy (the proportion of true negatives). An ideal prediction test has sensitivity and specificity equal to 100%. This means that all the healthy individuals are correctly classified as having low risk and all the diseased individuals are correctly classified as having high risk. For an imperfect test the sensitivity and specificity depend on the threshold T chosen and, therefore, describe prediction accuracy under a given threshold only. In this case, the Receiver Operating Characteristic (ROC) plot and the Area Under the ROC Curve (the AUC) is widely used. A ROC plot shows the sensitivity in relation to the $1 -$ specificity. Figure 1.3 illustrates the an example of ROC plot with four ROC curves. The AUC is a convenient characteristic of a binary prediction test that shows how much, on average, the sensitivity (proportion of true positives) exceeds the $1 -$ specificity (proportion of false positives) under different thresholds T. The minimum AUC value of 50% indicates no prediction power. Under such an AUC value the sensitivity (proportion of true positives) vs. the $1 -$ specificity (proportion of false positives) is 50%, which is simply equivalent to tossing a coin. The maximum AUC value of 100% indicates perfect prediction. Under such AUC value the sensitivity

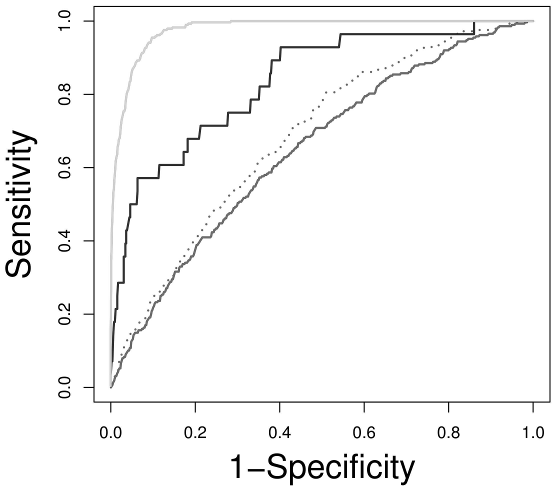


Figure 1.3: An example of an ROC plot. Accuracy to discriminate the top 5% tallest person, as measured by AUC, using different height profiles. (a) 54-loci genomic profile explaining 3.8% (54 loci, solid red line, AUC = 65% in the Rotterdam Study), population-specific 54-loci genomic profile explaining 5.8% in the Rotterdam Study (estimated using the data, red dotted line, $AUC = 68\%$ in the Rotterdam Study), mid-parental value explaining 40% (blue line, AUC=83% in the ERF study) and a hypothetical profile explaining 80% of height variance (green line, AUC97%). The plot taken from the Chapter 3.1, Figure 3.3, left panel.

(proportion of true positives) is 100% and the $1 -$ specificity (proportion of false positives) is 0% for any threshold T.

Height is a classical example of an inherited human trait. Typically, the proportion of the sex- and age-adjusted variance of height attributable to familial factors (the heritability) is estimated to be about 90% [26]. While height is among the most heritable human traits and many rare mutations lead to Mendelian diseases having very short stature as part of the syndrome, no genetic loci involved in the control of human height in the general population were known until recently. In 2008, three papers describing loci associated with height in the general population appeared in the 40th issue of Nature Genetics [27, 28, 29]. In total 54 loci showing strong statistical evidence for association with height were reported which all together explain approximately 5% of the variance in height. In the most recent GWAS on height [30] 183,727 individuals were used in the discovery of 180 loci influencing adult height. Many of these loci demonstrated patterns similar to what is also observed in other complex traits and diseases: those loci are connected with each other in biological pathways. It was shown that the causal genes are likely located near the most strongly associated variants. Moreover, many loci have multiple independently associated variants and many of which are involved in altering the amino acid structure of proteins or expression levels of nearby genes. The variants which exceeded the genome-wide significance threshold in this study explain $\approx 10\%$ of the total variance in height. However, inclusion of variants which did not reach this threshold increase the amount of variance explained up to $\approx 20\%$. Future even larger meta GWAS on height will likely increase this value. However, results of a study [31] suggest that the total variance explained by common SNPs is limited to approximately 45%.

More than 100 years ago, Francis Galton used height data to study the resemblance between parents and offspring, concluding that "when dealing with the transmission of stature from parents to children, the average height of the two parents, (. . .) is all we need care to know about them" [32]. In Chapter 3.1, I investigate the predictive potential of genomic profiling for complex human traits, and, as an instructive example, compare it to the 122-year-old Victorian method of Galton. We show that while a genomic profile based on 54 loci identified by 2008 explains only about 5% of height variance, the Victorian method of Galton can explain almost 40%.

Studying extreme phenotypes

Besides GWAS, there are many possible future strategies for studying a complex architecture of common traits. We addressed some of them in the previous section (i.e., compound heterozygote LOF alleles, genetic variants interacting with each other and with environmental factors). Many of them showed a great potential in discovering new variants. However, for the future studies it is important to choose a range of strategies allowing to identify the genetic variants and patterns in genetic variation explaining the highest proportion of heritability. Among many researchers there is a common opinion that GWAS with larger sample sizes can be an effective strategy which will allow discovering many new genetic variants with smaller effects. There are many discussions about influence of rare variants with relatively high effects which are difficult to detect in GWAS and in linkage analysis. It can be shown through simulations for a hypothetical trait which is similar to type 2 diabetes that dozens of such variants would account for most familial aggregation [19]. A real example of such a variant is in the LDLR locus showing effect which causes hypercholesterolemia [33]. Johansen et al. [34] demonstrates accumulation of rare variants in GWAS-identified genes in individuals with hypertriglyceridemia.

Classically, the individuals with extreme total cholesterol levels have been used successfully to find rare variants with relatively large effect sizes. Substantial enrichness of extremes with such rare variants should be manifested through decreased discriminative (predictive) ability of common genetic variants and environmental factors in extreme levels. In Chapter 3.2 I studied the extent to which known common variants influence extreme levels of total cholesterol in two populations (i.e., the family-based Erasmus Rucphen Family study and the population-based Rotterdam Study). By a measurement of the Area Under the Curve (AUC) I examined ability of the 52 common genetic variants and 6 environmental factors known to be responsible for total cholesterol to discriminate extreme levels.

Using dense genotyping for common studying phenotypes

Despite the great successes of GWAS in the identification of SNPs related to various traits, many variants have been identified for a single trait but have never been investigated in relation to a related trait. While blood pressure plays a crucial role in stroke and many genes are known to be implicated in blood pressure regulation, therefore, those genes are candidates for association with stroke. In the context of type 2 diabetes, coronary artery disease and myocardial infarction, and quantitative traits related to these disease, the Metabochip was designed [35], which allows cheap genotyping of SNPs involved in certain diseases in large(r) case series. SNPs in linkage disequilibrium with the GWAS SNPs have been added to evaluate additional (causal) variants. The Metabochip is a custom Illumina array which allows fine-mapping of 257 loci previously associated in GWAS of 23 traits. It assays 196,725 SNP markers that includes 63,450 replication SNPs selected to follow up previously identified GWAS variants and 122,241 SNPs located in the loci harboring those variants. Consequently, the Metabochip is a cost-effective alternative to sequencing which facilitates zooming in into loci associated with metabolic traits.

Carotid artery stenosis is an important cause of stroke and given its metabolic background the Metabochip is a suitable tool to study its genetic background. The carotid artery is the large artery supplying the brain and the face with blood. An atheromatous plaque can narrow the inner surface of the carotid artery that leads to a restricted blood flow and subsequently an ischemic stroke or a transient ischemic attack (neuralgic dysfunction with symptoms such as temporary loss of vision, difficulty in speaking and so on). We used the Metabochip to study the genetic architecture of patients with recent ischemic stroke or transient ischemic attack in Chapter 3.3.

SCOPE OF THE THESIS

The genetic architecture underlying a given phenotype can be relatively simple in the case of a trait that is controlled by one or several single genes harboring few mutations each having a large effect on the trait. Because of their large effects it is relatively easy to detect such mutations in linkage analysis or in GWAS with small sample size and to date many of these loci have been discovered. Linkage analysis is a well known approach which, during the 90's, allowed revealing more then a thousand new disease/trait related genomic regions. GWAS is a relatively new approach which, however, during the last five years, substantially improved our understanding of genetic basis of many traits and diseases. In Chapter 2.1, the methodology and software tool `ProbABEL` is described which can facilitate discovering new genetic variants in GWAS. From the beginning, I participated in the process of developing the tool. I essentially contributed in the implementation of various features such as two-step mixed model based procedures, interaction testing and bringing the tool to the final user-friendly stage.

Despite to successfully revealing many new genetic variants in GWASs, those variants explain a proportion of heritability which for many traits is much smaller then it was expected under the *common disease / common variant* model. This rose a question about the extent to which this model can describe the genetic basis of common traits (missing heritability issue) and called for testing other genetic models which will allow discovering new genetic variants. There are a number of models proposed. In this thesis I was focused on a few of them. It has been proposed that heterozygous loss-of-function variants may account for the essential proportion of heritability. In Chapter 2.2, a novel method of testing for association between loss-of-function in the compound heterozygous state and a trait is described. I participated in data analysis and development of software. In particular, I implemented the method in a computationally efficient manner and integrated it in the GenABEL package. Another approach to dissect missing heritability is testing genetic variants on presence of interaction between them or environmental factors. In Chapter 2.3, I describe a novel method for discovering potentially interacting loci in genome-wide scans. I proposed a new statistical method and studied its properties. This method was independently studied and applied in other research groups and demonstrated a high potential in revealing new genetic variants. In Chapter 2.4, it is described how I improved the method and developed an appropriate software tool.

Prediction is an important application of GWAS findings. In the future it allows for a personal medical treatment assigned according to the genetic profile of a specific person. In Chapter 3.1, the study of predictive power of known common genetic variants associated with height is described. I participated in developing of analysis plan, conducting the analysis and interpretation of the results.

The missing heritability issue is still under discussion and the interesting question which it opened is how much of the undiscovered trait's genetic component is attributed to common genetic variants with relatively small effects and how much to rare variants with relatively large effects. There are multiple examples of rare variants with relatively large effects. In the case if those variants are responsible on a substantial part of heritability this will refocus attention from GWAS (which was designed to study common variants) to another statistical methods. In Chapter 3.2 I studied the extent to which the common variants influence extreme levels of total cholesterol that can improve our understanding of enrichness of extremes with rare variants with relatively large effects..

There is evidence that causal genetic variants are located close to common variants showing GWAS signal. Those variants can be uncovered in a sequencing studies, however, this is an expensive approach. The Metabochip can serve as a cheap and fast alternative to this. The Metabochip allows targeted high-dense genotyping of loci where common variants associated with metabolic traits are located. Subsequent GWAS of these variants can reveal new associations. In Chapter 3.3, GWAS of carotid artery stenosis using Metabochip was conducted. I contributed to the writing of the analysis plan, to the data analysis and to the interpretation of the results.

The goal of the research described in this thesis which the author has contributed essentially to is the development and application of the novel methods, approaches and computational tools which can facilitate studying complex genetic architecture of common traits.

BIBLIOGRAPHY

1. Botstein, D, White, R. L., Skolnick, M & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32,** 314–331 (May 1980).
2. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics* **33,** 228–237 (Mar. 2003).
3. Roberts, R. J. Restriction endonucleases. *CRC Critical Reviews in Biochemistry* **4,** 123–164 (Nov. 1976).
4. Mullis, K. B. Target amplification for DNA analysis by the polymerase chain reaction. *Annales De Biologie Clinique* **48,** 579–582 (1990).
5. Hearne, C. M., Ghosh, S & Todd, J. A. Microsatellites for linkage analysis of genetic traits. *Trends in Genetics: TIG* **8,** 288–294 (Aug. 1992).
6. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (Sept. 1996).
7. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 2001).
8. Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291,** 1304–1351 (Feb. 2001).
9. Consortium, T. I. H. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (Oct. 2005).
10. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (Oct. 2007).
11. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (Sept. 2010).
12. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (Nov. 2012).
13. Matsuzaki, H. *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nature Methods* **1,** 109–111 (Nov. 2004).
14. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)* **308,** 385–389 (Apr. 2005).
15. *HuGENavigator|GWAS Integrator|Search* <http://hugenavigator.net/HuGENavigator/gWAHitStartPage.do> (visited on 31/01/2013).

16. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5,** e1000529 (June 2009).

17. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* **84,** 210–223 (Feb. 2009).

18. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34,** 816–834 (2010).

19. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

20. McQuillan, R. *et al.* Evidence of inbreeding depression on human height. *PLoS genetics* **8,** e1002655 (2012).

21. Vazquez, C *et al.* Thirteen cystic fibrosis patients, 12 compound heterozygous and one homozygous for the missense mutation G85E: a pancreatic sufficiency/insufficiency mutation with variable clinical presentation. *Journal of medical genetics* **33,** 820–822 (Oct. 1996).

22. Singleton, A. B., Hardy, J., Traynor, B. J. & Houlden, H. Towards a complete resolution of the genetic architecture of disease. *Trends in genetics: TIG* **26,** 438–442 (Oct. 2010).

23. Mukherjee, B., Ahn, J., Gruber, S. B. & Chatterjee, N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American Journal of Epidemiology* **175,** 177–190 (Feb. 2012).

24. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet* **6,** e1000981 (2010).

25. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490,** 267–272 (Oct. 2012).

26. Visscher, P. M. *et al.* Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American journal of human genetics* **81,** 1104–1110 (Nov. 2007).

27. MN, W., H, L. & CM, L. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40,** 575–583 (2008).

28. G, L., AU, J. & C, G. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40,** 584–591 (2008).

29. DF, G., GB, W. & G, T. Many sequence variants affecting diversity of adult human height. *Nat Genet* **40,** 609–615 (2008).

30. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (Oct. 2010).

31. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42,** 565–569 (July 2010).

32. F, G. Regression towards mediocrity in hereditary stature. *Journal of the anthropological institute* **15,** 246–263 (1886).

33. Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS genetics* **7,** e1002198 (July 2011).

34. Johansen, C. T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature genetics* **42,** 684–687 (Aug. 2010).

35. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics* **8,** e1002793 (Aug. 2012).

36. 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).

37. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10,** 387–406 (2009).

38. Chen, W.-M. & Abecasis, G. R. Family-based association tests for genomewide association scans. *Am J Hum Genet* **81,** 913–926 (2007).

39. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).

40. Amin, N., van Duijn, C. M. & Aulchenko, Y. S. A genomic background based method for association analysis in related individuals. *PLoS One* **2,** e1274 (2007).

41. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nature genetics* **44,** 1166–1170 (Oct. 2012).

42. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30,** 97–101 (2002).

43. Perez-Enciso, M. & Misztal, I. Qxpak: a versatile mixed model application for genetical genomics and QTL analyses. *Bioinformatics* **20,** 2792–2798 (2004).

44. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42,** 348–354 (Apr. 2010).

45. *GenABEL.org GenABLE'ing genetical research* http://www.genabel.org/. <`http://www.genabel.org/`> (visited on 18/12/2012).

46. Woodward, O. M. *et al.* Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *Proc Natl Acad Sci U S A* **106,** 10338–10342 (2009).

47. 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).

48. Vink, J. M. *et al.* Genome-wide association study of smoking initiation and current smoking. *Am J Hum Genet* **84,** 367–379 (2009).

49. Estrada, K. *et al.* A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. *Hum Mol Genet* **18,** 3516–3524 (2009).

50. Rönnegård, L. & Valdar, W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC genetics* **13,** 63 (2012).

51. Deng, W. Q. & Paré, G. A fast algorithm to optimize SNP prioritization for gene-gene and gene-environment interactions. *Genetic epidemiology* **35,** 729–738 (Nov. 2011).

52. Sorensen, D. Developments in statistical analysis in quantitative genetics. *Genetica* **136,** 319–332 (June 2009).

53. Rönnegård, L. & Valdar, W. Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics* **188,** 435–447 (June 2011).

54. Visscher, P. M. & Posthuma, D. Statistical power to detect genetic Loci affecting environmental sensitivity. *Behavior genetics* **40,** 728–733 (Sept. 2010).

55. AC, J., MC, P., EW, S. & van Duijn CM. Revisiting the clinical validity of multiplex genetic testing in complex diseases. *Am J Hum Genet* **74,** 585–588 (2004).

56. AC, J. *et al.* Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* **8,** 395–400 (2006).

57. Demirkan, A. *et al.* Genetic architecture of circulating lipid levels. *European journal of human genetics: EJHG* **19,** 813–819 (July 2011).

58. Demirkan, A *et al.* Genetic risk profiles for depression and anxiety in adult and elderly cohorts. *Molecular Psychiatry* **16,** 773–783 (2011).

59. Gibson, G. Rare and common variants: twenty arguments. *Nature reviews. Genetics* **13,** 135–145 (Feb. 2011).

60. Chan, Y. *et al.* Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. *PLoS genetics* **7,** e1002439 (Dec. 2011).

61. Jiang, T., Yang, L., Jiang, H., Tian, G. & Zhang, X. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. *Science China. Life sciences* **54,** 945–952 (Oct. 2011).

62. Keating, B. J. *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PloS one* **3,** e3583 (2008).

63. Kang, H. S. *et al.* Transcription factor Glis3, a novel critical player in the regulation of pancreatic beta-cell development and insulin gene expression. *Molecular and cellular biology* **29,** 6366–6379 (Dec. 2009).

64. Boesgaard, T. W. *et al.* Variants at `DGKB/TMEM195`, `ADRA2A`, `GLIS3` and `C2CD4B` loci are associated with reduced glucose-stimulated beta cell function in middle-aged Danish people. *Diabetologia* **53,** 1647–1655 (Aug. 2010).

65. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics* **42,** 105–116 (Feb. 2010).

66. Hu, C. *et al.* Variants from `GIPR`, `TCF7L2`, `DGKB`, `MADD`, `CRY2`, `GLIS3`, `PROX1`, `SLC30A8` and `IGF1` are associated with glucose metabolism in the Chinese. *PloS one* **5,** e15542 (2010).

67. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics* **41,** 703–707 (June 2009).

68. Göksan, B, Erkol, G, Bozluolcay, M & Ince, B. Diabetes as a determinant of high-grade carotid artery stenosis: evaluation of 1,058 cases by Doppler sonography. *Journal of stroke and cerebrovascular diseases: the official journal of National Stroke Association* **10,** 252–256 (Dec. 2001).

69. Inchiostro, S *et al.* Prevalence of diabetes and/or ischaemic heart disease in classes of increasing carotid artery atherosclerosis: an ultrasonographic study. *Diabetic medicine: a journal of the British Diabetic Association* **20,** 670–676 (Aug. 2003).

70. Folsom, A. R. *et al.* Relation of carotid artery wall thickness to diabetes mellitus, fasting glucose and insulin, body size, and physical activity. Atherosclerosis Risk in Communities (ARIC) Study Investigators. *Stroke; a journal of cerebral circulation* **25,** 66–73 (Jan. 1994).

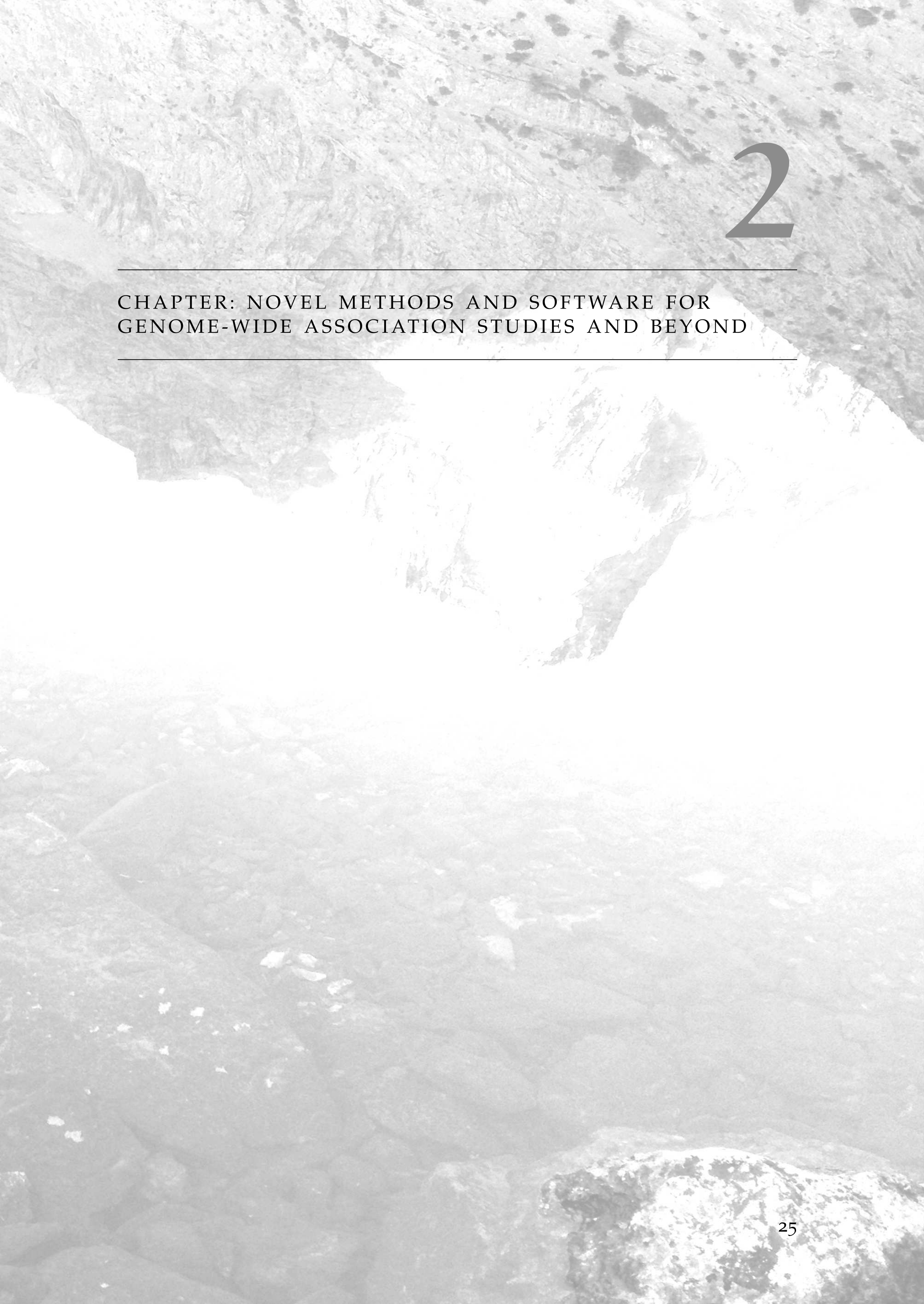
71. Tropeano, A., Boutouyrie, P., Katsahian, S., Laloux, B. & Laurent, S. Glucose level is a major determinant of carotid intima-media thickness in patients with hypertension and hyperglycemia. *Journal of hypertension* **22,** 2153–2160 (Nov. 2004).

72. *ROOT A Data Analysis Framework* http://root.cern.ch/drupal/. `<http://root.cern.ch/drupal/>` (visited on 17/12/2012).

73. *DatABEL package GenABEL.org* http://www.genabel.org/packages/DatABEL. `<http://www.genabel.org/packages/DatABEL>` (visited on 17/12/2012).

74. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90,** 7–24 (Jan. 2012).

75. Goldstein, D. B. Common genetic variation and human traits. *The New England journal of medicine* **360,** 1696–1698 (Apr. 2009).



2

CHAPTER: NOVEL METHODS AND SOFTWARE FOR GENOME-WIDE ASSOCIATION STUDIES AND BEYOND

2.1 probabel PACKAGE FOR GENOME-WIDE ASSOCIATION ANALYSIS OF IMPUTED DATA

Yurii S Aulchenko[1,2]***, Struchalin M***[1] ***and Cornelia M van Duijn***[1]

[1] *Department of Epidemiology, Erasmus MC, Postbus 2040, 3000 CA Rotterdam, The Netherlands*
[2] *Institute of Cytology and Genetics SD RAS, Novosibirsk, 630090, Russia*

Abstract

BACKGROUND: Over the last few years, genome-wide association (GWA) studies became a tool of choice for the identification of loci associated with complex traits. Currently, imputed single nucleotide polymorphisms (SNP) data are frequently used in GWA analyzes. Correct analysis of imputed data calls for the implementation of specific methods which take genotype imputation uncertainty into account.
RESULTS: We developed the ProbABEL software package for the analysis of genome-wide imputed SNP data and quantitative, binary, and time-till-event outcomes under linear, logistic, and Cox proportional hazards models, respectively. For quantitative traits, the package also implements a fast two-step mixed model-based score test for association in samples with differential relationships, facilitating analysis in family-based studies, studies performed in human genetically isolated populations and outbred animal populations.
CONCLUSIONS: ProbABEL package provides fast efficient way to analyze imputed data in genome-wide context and will facilitate future identification of complex trait loci.

BMC Bioinformatics,
2010 Mar 16;11:134

Genome-wide association (GWA) studies became the tool of choice for the identification of loci associated with complex traits. In GWA analyses, association between a trait of interest and genetic polymorphisms (usually single nucleotide polymorphisms, SNPs) is studied using thousands of people typed for hundreds of thousands of polymorphisms. Several hundred loci for dozens of complex human disease and quantitative traits have been discovered thus far using this method [1].

For any given genetic polymorphism, association can be studied using standard statistical analysis methodology, such as fixed and mixed effects models. However, because of the large number of tests to be performed and the quantity of data to be stored in GWA studies, computational throughput and effective data handling are essential features of statistical analysis software to be used in this context. A number of specialized software packages, such as `PLINK` [2], `GenABEL` [3], `SNPTEST` [4] and `snpMatrix` [5] were developed for the statistical analysis of GWA data. Most of these packages were designed, and are fit for, the analysis of directly typed SNPs. When directly typed markers are studied, genotype calling is performed with a high degree of confidence for the vast majority of markers, resulting in four possible genotypes ("AA", "AB", "BB", and missing). This allows representation of each individual genotype using two-bit coding and consequently effective storage of the genotype data in RAM [3].

Recently, novel statistical tools for genotype imputations [6, 4, 7, 8, 9] and experimental techniques for high-throughput sequencing were developed. Implementation of these methods usually results in estimates of the posterior probability distributions $\mathbf{P}_g = (P_{AA}, P_{AB}, P_{BB})$ of the genotypes based on the available data. For many genomic loci, this distribution may be non-degenerate.

Several techniques can be applied to analysis of such "uncertain" data. The most simplistic approach would be to use the "best guess genotypes", that is to use the genotype with the highest posterior probability ($g = \max_g P_g$) for analysis as if it were a directly typed markers. This approach is equivalent to replacing the estimated probability distribution with a degenerate one where a probability of one is assigned to the genotype with the maximal posterior probability. From standard statistical theory it is known, however, that such a procedure results in biased estimates of the effects. A correct analysis can be achieved using a maximum likelihood approach. Under this approach the likelihood can be computed using the total probability formula in which summation is performed over the genotypes, whose true values are not known, but whose posterior probabilities can be estimated given the data. This approach is computationally demanding, as it requires summation over the underlying probability distribution and numerical maximization of the likelihood function. Alternatively, a regression approach in which the posterior genotypic probabilities are used as predictors, can be applied. The main advantage of this approach is that well-established regression analysis methodology, algorithms, and code can be used in its implementation. Most currently available packages for GWA analysis can not be directly used in this manner, as they assume degenerate genotypic distributions and do not provide a facility for the storage and analysis of real-number predictors (posterior genotypic probabilities).

In this work, we describe the ProbABEL package, which was designed to perform genome-wide regression on posterior genotypic probabilities estimated using imputation software, such as MACH [6] or IMPUTE [4, 9]. In addition to standard linear and logistic regression, which is widely applied to the analysis of quantitative and binary outcomes in population-based GWA studies, we also implemented a Cox proportional hazards model. For quantitative traits, we implemented a fast two-step mixed model-based score test for association testing in studies with a high degree of confounding induced by differential relationships between study subjects (e.g. family-based studies, studies of human genetically isolated populations, and studies in outbred animal populations).

Implementation

Here, in the first few sub-subsections, we will describe ProbABEL software, giving only the main outline of the underlying theory and with special emphasis on implementation and the options allowing to access specific analyzes within ProbABEL. In two last sub-subsections, starting with the "Fixed effects model theory", we will give more in-depth review of the theory used by the package.

ProbABEL was implemented using code written in the C and C++ languages. The package consists of three executable files, used to perform linear, logistic, and Cox regressions, and a helper Perl script which facilitates the analysis of multiple chromosomes.

The package implements standard regression analysis methodology outlined in the subsection "Fixed effects model theory" and specific approximation to the mixed linear model described in the subsection "Two-step score test approximation to the mixed model". The key statistical tests performed by ProbABEL concern testing of the SNP effects. Here, we will describe the tests performed by ProbABEL using an example of linear regression; testing using other types of regression follows similar logic.

In linear regression, the expectation of the trait is described as

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_x\boldsymbol{\beta}_x + \mathbf{X}_g\boldsymbol{\beta}_g,$$

where $\mathbf{Y}$ is the vector of phenotypic values, $\mathbf{X}_g$ is the design matrix containing data about predictors of interest (these involving SNP data), and $\mathbf{X}_x$ is the design matrix containing other (nuisance) covariates. $\boldsymbol{\beta}_g$ and $\boldsymbol{\beta}_x$ are the vectors of corresponding fixed effects. The vector of phenotypes $\mathbf{Y}$ and the covariates matrix $\mathbf{X}_x$ are provided in the phenotype file. The genotypic data are read from the genotype (dose or probability) files and are analyzed one SNP at a time.

Our interest lies in testing the (components of) $\boldsymbol{\beta}_g$. ProbABEL provides the estimates of the components of the vector $\boldsymbol{\beta}_g$ and corresponding standard errors, and, in most cases, the test of the general hypothesis concerning the involvment of the SNP, obtained by comparison of the estimated model to the null model formulated as $\boldsymbol{\beta}_{g,0} = \mathbf{0}$, where $\mathbf{0}$ is the vector of zeros.

Under the general genotypic model, $\mathbf{X}_g$ is a matrix with the number of rows equal to the number of people under consideration and with two columns. Each row of the matrix contains the estimated probabilities that a person has genotype "AA" or "AB". Then, the vector of genotypic effects is described with two parameters: $\boldsymbol{\beta}_g = (\beta_{AA}, \beta_{AB})$. Thus formulated, the model allows for the estimation of a general genotypic two-degree of freedom model. Further, a number of sub-models can be formulated by setting restrictions on these parameters. The "dominant B allele" model is formalized as $\beta_{AB} = 0$, "dominant A" (the same as "recessive B") as $\beta_{AA} = \beta_{AB}$, the additive model as $2 \cdot \beta_{AB} = \beta_{AA}$, and the over-dominant model as $\beta_{AA} = 0$. Note that the additive model is equivalent to performing linear regression on the estimated dose of allele "A" defined as $P_{AB} + 2 \cdot P_{AA}$. The latter model is tested when the allelic dosage file is provided as the input for ProbABEL, while the full range of described models is tested if the estimated probability files (option "-ngpreds=2") are supplied.

ProbABEL can also test for interaction between a specified covariate and the set of SNPs; for that alternative, the interaction covariate should be specified using the "-interaction N" option, where N corresponds to the number of the column of the design matrix $\mathbf{X}_x$, which contains that covariate. If this option is used, the expectation of the trait is defined as

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_x\boldsymbol{\beta}_x + \mathbf{X}_g\boldsymbol{\beta}_g + (\mathbf{X}_g^T\mathbf{W})^T\boldsymbol{\beta}_{\mathbf{gxe}},$$

where $\mathbf{W}$ is a diagonal matrix, whose diagonal elements are formed by substituting the interaction covariate to the matrix and $\boldsymbol{\beta}_{gxe}$ is the vector of interaction regression coefficients.

Analysis of population-based data

If the study subjects can be assumed to be genetically "independent", in the sense that they come from the general outbred population without a marked degree of stratification and that cryptic relatedness is absent, the data can be effectively analyzed using standard linear fixed effects regression methodology, as described in subsection "Fixed effects model theory". The (small) effects of confounding can be corrected posterior to analysis using the genomic control [10] procedure. If a marked degree of stratification is present, such methods as structured association analysis and EIGENSTRAT [11] can be combined with the standard methods.

Using standard methods, the estimates of the parameters can be obtained using the standard formula 2.1 (see "Fixed effects model theory" below), which provides maximum likelihood estimates if $(X^TX)^{-1}$ exists. The latter condition is fulfilled for virtually all analyses; practically, exceptions may occur for SNPs with very low minor allele frequencies or poor quality imputations.

The standard errors are computed as square roots of the diagonal elements of the parameter estimates' variance-covariance matrix. This matrix is computed using one of three different methods: the standard method, with residual variance estimated under the alternative (formula 2.2, see "Fixed effects model theory" below) or null hypothesis concerning SNPs (option "`-score`"), or using a "sandwich" estimator (formula 2.5, see "Fixed effects model theory"), resulting in robust standard errors (option "`-robust`").

The value of the global likelihood ratio test statistic, testing the joint significance of all terms involving SNP, is computed using the formula 2.3 (see "Fixed effects model theory"). In this test, the null model is formulated as $\boldsymbol{\beta}_{g,0} = \mathbf{0}$, where $\mathbf{0}$ is the vector of zeros. If an interaction term is present, that is also set to zero under the null: $\boldsymbol{\beta}_{gxe,0} = \mathbf{0}$. The likelihoods involved are computed using the formula 2.4 (see "Fixed effects model theory") with the values of the parameters fixed at the point of the maximum likelihood estimate obtained with 2.1 (see "Fixed effects model theory").

Analysis of data on subjects with differential relationships

In the case of a study involving subjects with markedly differential relationships (family-based designs, studies of human genetically isolated populations, studies in outbred animal populations), a mixed model approach may be used, in which a random effect ("heritability") accounts for similarities between the phenotypes of study subjects [12]. However, the estimation of the full mixed model using either maximum likelihood or the restricted maximum likelihood approach is computationally demanding, if not unfeasible, within the framework of GWAS [13], and therefore a two-step mixed model-based approach [13, 14, 15] is utilized in `ProbABEL`.

In this approach, the mixed model containing all terms but those involving SNP is first estimated by maximizing the likelihood function provided by the expression 2.7 (see subsection "Two-step score test approximation to the mixed model" for details). These estimates are then used in the second step to compute estimates of the SNP effects (formula 2.8 of "Two-step score test approximation to the mixed model") and the variance-covariance matrix of these estimates (formula 2.10, see "Two-step score test approximation to the mixed model"). These values can be used to perform a score test for association.

The second step of a mixed-model based score test for association is available in `ProbABEL` using option "`-mmscore IVFile`", where `IVFile` is the name of a file containing the inverse of the variance-covariance matrix ($\mathbf{V}^{-1}_{\hat{h}^2,\hat{\sigma}^2}$ of formulas 2.8 and 2.10, see "Two-step score test approximation to the mixed model") evaluated at the point of the maximum likelihood estimates obtained in step one. The phenotypes analyzed in the second step are residuals (as specified by the formula 2.9, see "Two-step score test approximation to the mixed model") obtained by subtracting the trait values expected under the mixed model-based estimates of the fixed effects from the original trait values.

Step one of the regression procedure can be performed using our `GenABEL` software [3]. This software performs genomic data based estimation of the kinship matrix as described in subsection ″Estimation of genomic kinship matrix″ using the `ibs(...,weight="freq")` function, and performs maximum likelihood estimation of the step-one mixed model using the `polygenic()` function. The resulting object contains the inverse variance-covariance matrix (`object$InvSigma`), which can be saved as a text file and used in `ProbABEL` analysis. The residuals to be used as trait values in step two of the analysis can be accessed through `object$residualY`.

Input and output

The input consists of a phenotypic data file and a set of files describing the imputed genotypic data. The phenotypic file provides data on the outcome of interest and any additional covariates to be included in the analysis. The genotypic data files, at present, utilize the MACH imputation software output format. Minimally, a file with estimated probability distributions (″mlprob″) or allelic dosages (″mldose″) and the ″mlinfo″ file containing information about allele coding and overall imputation quality should be provided. Optionally, a map file in HapMap format, containing chromosome and location information, may be supplied. Information contained in the latter two files is not used in analysis, but is forwarded directly to the output. If the mixed-model based score test for association in related individuals is to be computed, a file containing the inverse matrix of variances and covariances between the phenotypes of study individuals should be supplied as a part of the input. The output of the program consists of one line for each SNP tested, containing information about the SNP supplied as part of the input, as well as the results from analysis (estimates of the coefficients of regression, standard errors of the coefficients, and test statistic values).

Fixed effects model theory

Most of the fixed effects model theory outlined here is standard and can be found in textbooks, such as ″Generalized, Linear, and Mixed Models″ [16]. Specific references are provided when this is not the case.

LINEAR REGRESSION ASSUMING NORMAL DISTRIBUTION Standard linear regression theory is used to estimate coefficients of regression and their standard errors. We assume linear model with expectation

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

and variance-covariance matrix

$$\mathbf{V} = \sigma^2 \mathbf{I},$$

where $\mathbf{Y}$ is the vector of phenotypes of interest, $\mathbf{X}$ is design matrix, $\boldsymbol{\beta}$ is the vector of regression parameters, σ^2 is variance and $\mathbf{I}$ is identity matrix.

The maximum likelihood estimates (MLEs) for the regression parameters is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \tag{2.1}$$

and MLE of the residual variance is

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{N - r_X},$$

where N is the number of observations and r_X is rank of $\mathbf{X}$ (number of columns of the design matrix).

The variance-covariance matrix for the parameter estimates under alternative hypothesis can be computed as

$$\mathbf{var}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}. \tag{2.2}$$

For the j-the element $\hat{\boldsymbol{\beta}}(j)$ of the vector of estimates the standard error under alternative hypothesis is given by the square root of the corresponding diagonal element of the above matrix, $\mathbf{var}_{\hat{\boldsymbol{\beta}}}(jj)$, and the Wald test can be computed with

$$T^2(j) = \frac{\hat{\boldsymbol{\beta}}(j)^2}{\mathbf{var}_{\hat{\boldsymbol{\beta}}}(jj)},$$

which asymptotically follows the χ^2 distribution with one degree of freedom under the null hypothesis.

When testing significance for more than one parameter simultaneously, several alternatives are available. Let us partition the vector of parameters into two components, $\boldsymbol{\beta} = (\boldsymbol{\beta}_g, \boldsymbol{\beta}_x)$, and our interest is testing the parameters contained in $\boldsymbol{\beta}_g$ (SNP effects), while $\boldsymbol{\beta}_x$ (e.g. effects of sex, age, etc.) are considered nuisance parameters. Let us define the vector of the parameters of interest which are fixed to certain values under the null hypothesis as $\boldsymbol{\beta}_{g,0}$ (usually, $\boldsymbol{\beta}_{g,0} = \mathbf{0}$, vector of zeros).

The likelihood ratio test can be obtained with

$$LRT = 2 \cdot (logLik(\hat{\boldsymbol{\beta}}_g, \hat{\boldsymbol{\beta}}_x) - logLik(\boldsymbol{\beta}_{g,0}, \hat{\boldsymbol{\beta}}_x)), \tag{2.3}$$

which under the null hypothesis is asymptotically distributed as χ^2 with number of degrees of freedom equal to the number of parameters specified by $\boldsymbol{\beta}_g$. Assuming the normal distribution, the log-likelihood of a model specified by the vector of parameters $\boldsymbol{\beta}$ and residual variance σ^2 can be computed as

$$logLik(\boldsymbol{\beta},\sigma^2) = -\frac{1}{2}(N \cdot log_e\sigma^2 + (\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})^T(\mathbf{I}/\sigma^2)(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})). \tag{2.4}$$

Secondly, the Wald test can be used; for that the inverse variance-covariance matrix of $\hat{\boldsymbol{\beta}}_g$ should be computed as

$$\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}_g} = \mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(g,g) - \mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(g,x)(\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(x,x))^{-1}\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(x,g),$$

where $\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(a,b)$ correspond to sub-matrices of the inverse of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, involving either only covariances between the parameters of interest (g,g), only the nuisance parameters (x,x) or between the parameters of interest and nuisance parameters, (x,g), (g,x).

The Wald test statistics is then computed as

$$W^2 = (\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0})^T\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}_g}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0}),$$

which asymptotically follows the χ^2 distribution with the number of degrees of freedom equal to the number of parameters specified by $\boldsymbol{\beta}_g$. The Wald test generally is computationally easier than the LRT, because it avoids estimation of the model specified by the parameter's vector $(\boldsymbol{\beta}_{g,0}, \hat{\boldsymbol{\beta}}_x)$.

Lastly, similar to the Wald test, the score test can be performed by use of $\mathbf{var}_{\boldsymbol{\beta}=(\boldsymbol{\beta}_{g,0},\hat{\boldsymbol{\beta}}_x)}$ instead of $\mathbf{var}_{\hat{\boldsymbol{\beta}}}$.

LOGISTIC REGRESSION For logistic regression, the procedure to obtain parameters estimates, their variance-covariance matrix, and tests are similar to these outlined above with several modifications.

The expectation of the binary trait is defined as expected probability of the event as defined by the logistic function

$$E[\mathbf{Y}] = \pi = \frac{1}{1 + e^{-(\mathbf{X}\boldsymbol{\beta})}}.$$

The estimates of the parameters are obtained not in one step, as is the case of the linear model, but using iterative procedure (iteratively re-weighted least squares). This procedure is not described here for the sake of brevity.

The log-likelihood of the data is computed using binomial probability formula:

$$logLik(\boldsymbol{\beta}) = \mathbf{Y}^T log_e\pi + (\mathbf{1} - \mathbf{Y})^T log_e(\mathbf{1} - \pi),$$

where $log_e\pi$ is a vector obtained by taking the natural logarithm of every value contained in the vector π.

ROBUST VARIANCE-COVARIANCE MATRIX OF PARAMETER ESTIMATES For computations of robust variance-covariance matrix we use White's sandwich estimator [17, 18], which is equivalent to the "HC0" estimator described by Zeilers and Lumley in "sandwich" package for R.

For linear model, the variance-covariance matrix of parameter estimates is computed using formula

$$\mathbf{var}_r = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{R}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}, \tag{2.5}$$

where $\mathbf{R}$ is a diagonal matrix containing squares of residuals of $\mathbf{Y}$. The same formula may be used for "standard" analysis, in which case the elements of the $\mathbf{R}$ matrix are constant, namely mean residual sum of squares (the estimate of residual variance, $\hat{\sigma}^2$).

Similar to that, the robust matrix is computed for logistic regression with

$$\mathbf{var}_r = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{R}\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1},$$

where $\mathbf{W}$ is the diagonal matrix of "weights" used in logistic regression.

COX PROPORTIONAL HAZARDS MODEL The implementation of the Cox proportional hazard model used in ProbABEL is entirely based on the code of R library survival developed by Thomas Lumley (function coxfit2), and is therefore not described here.

Two-step score test approximation to the mixed model

The framework for analysis of data containing differential relationships follows the two-step logic developed in the works of Aulchenko et al. [3] and Chen and Abecasis [14]. General analysis model is a linear mixed model which defines the expectation of the trait as

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

identical to that defined for linear model. To account for possible correlations between the phenotypes of study subjects the variance-covariance matrix is defined to be proportional to the linear combination of the identity matrix $\mathbf{I}$ and the relationship matrix $\mathbf{\Phi}$:

$$\mathbf{V}_{\sigma^2,h^2} = \sigma^2(2h^2\mathbf{\Phi} + (1-h^2)\mathbf{I}),$$

where h^2 is the heritability of the trait. The relationship matrix $\mathbf{\Phi}$ is twice the matrix containing the coefficients of kinship between all pairs of individuals under consideration; its estimation is discussed in a separate subsection "Estimation of genomic kinship matrix".

Estimation of thus defined model is possible by numerical maximization of the likelihood function, however, the estimation of such model for large data sets is not computationally feasible for hundreds of thousands to millions of SNPs tested in the context of GWAS, as we have demonstrated previously [3].

TWO-STEP SCORE TEST FOR ASSOCIATION A two-step score test approach is therefore used to decrease the computational burden. Let us re-write the expectation of the trait by splitting the design matrix in two parts, the "base" part $\mathbf{X}_x$, which includes all terms not changing across all SNP models fit in GWAS (e.g. effects of sex, age, etc.), and the part including SNP information, $\mathbf{X_g}$:

$$E[\mathbf{Y}] = \mathbf{X}_x\boldsymbol{\beta}_x + \mathbf{X}_g\boldsymbol{\beta}_g. \tag{2.6}$$

Note that the latter design matrix may include not only the main SNP effect, but e.g. SNP by environment interaction terms.

At the first step, linear mixed model not including SNP effects

$$E[\mathbf{Y}] = \mathbf{X}_x\boldsymbol{\beta}_x$$

is fitted. The maximum likelihood estimates (MLEs) of the model parameters (regression coefficients for the fixed effects $\hat{\boldsymbol{\beta}}_x$, the residual variance $\hat{\sigma}^2_x$ and the heritability $\hat{h}^2_x$) can be obtained by numerical maximization of the likelihood function

$$logLik(\boldsymbol{\beta}_x, h^2, \sigma^2) = -\frac{1}{2}(log_e|\mathbf{V}_{\sigma^2,h^2}| + (\mathbf{Y} - \boldsymbol{\beta}_x\mathbf{X}_x)^T\mathbf{V}^{-1}_{\sigma^2,h^2}(\mathbf{Y} - \boldsymbol{\beta}_x\mathbf{X}_x)), \tag{2.7}$$

where $\mathbf{V}^{-1}_{\sigma^2,h^2}$ is the inverse and $|\mathbf{V}_{\sigma^2,h^2}|$ is the determinant of the variance-covariance matrix.

At the second step, the estimates of the fixed effects of the terms involving SNP are obtained with

$$\hat{\boldsymbol{\beta}}_g = (\mathbf{X}_g^T\mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2}\mathbf{X}_g)^{-1}\mathbf{X}_g^T\mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2}\mathbf{R}_{\hat{\beta}_x}, \tag{2.8}$$

where $\mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2}$ is the variance-covariance matrix at the point of the MLE estimates of $\hat{h}^2_x$ and $\hat{\sigma}^2_x$ and

$$\mathbf{R}_{\hat{\beta}_x} = \mathbf{Y} - \hat{\boldsymbol{\beta}}_x\mathbf{X}_x \tag{2.9}$$

is the vector of residuals obtained from the base regression model. Under the null model, the inverse variance-covariance matrix of the parameter's estimates is defined as

$$\mathbf{var}_{\hat{\beta}_g} = \hat{\sigma}^2_x(\mathbf{X}_g^T\mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2}\mathbf{X}_g)^{-1}. \tag{2.10}$$

Thus the score test for joint significance of the terms involving SNP can be obtained with

$$T^2 = (\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0})^T\mathbf{var}^{-1}_{\hat{\beta}_g}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0}),$$

where $\boldsymbol{\beta}_{g,0}$ are the values of parameters fixed under the null model. This test statistics under the null hypothesis asymptotically follows the χ^2 distribution with the number of degrees of freedom equal to the number of parameters tested. The significance of an individual j-the elements of the vector $\hat{\boldsymbol{\beta}}_g$ can be tested with

$$T_j^2 = \frac{\hat{\boldsymbol{\beta}}_g^2(j)}{\mathbf{var}_{\hat{\beta}_g}(jj)},$$

where $\hat{\boldsymbol{\beta}}_g^2(j)$ is square of the j-th element of the vector of estimates $\hat{\boldsymbol{\beta}}_g$, and $\mathbf{var}_{\hat{\beta}_g}(jj)$ corresponds to the j-th diagonal element of $\mathbf{var}_{\hat{\beta}_g}^{-1}$. This statistics asymptotically follows χ_1^2.

ESTIMATION OF GENOMIC KINSHIP MATRIX The relationship matrix $\boldsymbol{\Phi}$ used in estimation of the linear mixed model is twice the matrix containing the coefficients of kinship between all pairs of individuals under consideration. This coefficient is defined as the probability that two gametes randomly sampled from each member of the pair are identical-by-descent (IBD), that is they are copies of exactly the same ancestral allele. The expectation of kinship can be estimated from pedigree data using standard methods, for example the kinship for two outbred sibs is 1/4, for grandchild-grandparent is 1/8, etc. However, in many situations, pedigree information may be absent, incomplete, or not reliable. Moreover, the estimates obtained using pedigree data reflect the expectation of kinship, while the true realization of kinship may vary around this expectation. In presence of genomic data it may therefore be desirable to estimate the kinship coefficient from these, and not from pedigree. It can be demonstrated that unbiased and positive semi-definite estimator of the kinship matrix [19] can be obtained by computing the kinship coefficients between individuals i and j with

$$\hat{K}_{ij} = \frac{1}{L}\sum_{l=1}^{L}\frac{(g_{l,i}-p_l)(g_{l,j}-p_l)}{p_l(1-p_l)},$$

where L is the number of loci, p_l is the allelic frequency at l-th locus and $g_{l,j}$ is the genotype of j-th person at the l-th locus, coded as 0, 1/2, and 1, corresponding to the homozygous, heterozygous, and other type of homozygous genotype [15, 19, 11]. The frequency is computed for the allele which, when homozygous, corresponds to the genotype coded as "1'.

Results

To ensure the statistical correctness of the two-step procedure, we performed a small-scale simulation study. We used real data from the Erasmus Rucphen Family (ERF) study [20]. In simulations, we used genotypic data from 2,313 people who had high-density SNP genotyping data. The trait was simulated as a sum of four independent effects: two fixed effects explaining 10 and 5%

of the total trait variance, a polygenic effect, and a residual random effect. The residual random effect was assumed to be distributed normally with mean zero and variance fixed at the value that explained 59.5% of total variance. To simulate the polygenic effect, similar to our previous work [15], we selected 200 random SNPs, and assigned these SNPs with fixed effects such that, in total, these SNPs explained 25.5% of total variance. Thus, the heritability of the trait when adjusted for the fixed effects was 30%.

The SNPs mimicking the polygenic effect were selected randomly from all autosomes but the second. To estimate type 1 error of the two-step procedure, we studied association of the trait with the second chromosome SNPs using real imputed data. Only SNPs with estimated minor allele frequencies greater than 1% were used in analysis (212,691 SNPs in total). We compared type 1 error rates for four different models: a linear model ignoring the relatedness structure (using both a standard and a robust covariance matrix) and our 2-step mixed-model based score test. For the latter, we adjusted for two fixed effect covariates in the first step (polygenic) analysis.

The results of these tests are summarized in Table 1. It is easy to see that when relationships between study individuals are not taken into account, the distribution of the test statistic is inflated, regardless of whether a robust or standard covariance matrix is used. In our previous work, we demonstrated that this inflation grows with increasing trait heritability, with more close relatives present in the sample [15] and with increasing sample size and can reach very high values. On the contrary, when two-step approximation to the mixed model is used ("Linear, mmscore" row of Table 1), the test statistic shows very good agreement to the $\chi^2_{df=1}$ distribution expected under the null.

Next, we measured CPU time required for particular ProbABEL analyses. To do this, we selected 500, 1000, and 1500 people from 2,313 genotyped individuals and measured the speed of different types of analysis using chromosome 2 imputed data on 220,833 SNPs. All analyses were ran on a Sun Fire X4640 server with an Intel Xeon CPU 5160 (3.00 GHz). Results are present in Table 2. From this table, it is clear that all population-based analyzes (these not involving the `-mmscore` option) scale roughly linearly with the number of people. Use of the `-robust` option increases the running time by only a small fraction. Based on these data, one would expect that a GWA analysis involving, for example, 2.5 millions SNPs imputed on HapMap2 release 22 in 1,500 individuals would take 1/2 hour for linear, 2 hours for logistic and 1 1/2 hours for Cox proportional hazards models.

Use of the `-mmscore` option to adjust for relationships between study subjects, however, induces a non-linear relationship between the number of study subjects and analysis time: while the time to analyze 500 people is 16 minutes, the time for analysis of 1500 people is $\approx$ 14 times longer. The time for a GWA with 1,500 people and 2.5 millions imputed SNPs is, therefore, estimated to be $\approx$ 43 hours.

Imputed SNP data are conventionally used for the analysis of GWA data; correct use of imputed data allows for higher power and location accuracy [21, 22]. However, correct analysis of imputed data needs to account for the uncertainty surrounding estimated genotypic probability distributions. This can be done using approaches based on either likelihood or regression on estimated probabilities, as outlined in the ”Background” and ”Implementation”. A number of software packages are available for such analyses. SNPTEST implements a score test based on missing data likelihood [4] allowing for the study of both quantitative and binary outcomes. MACH2QTL and MACH2DAT implement regression models on estimated probabilities for quantitative and binary traits, respectively, in a manner similar to ProbABEL. ProbABEL extends the functionality available in these packages by allowing analysis under the Cox proportional hazards model. Further, while SNPTEST allows for testing interaction of a covariate with SNPs studied, it does not provide the value of the global significance test. Finally, ProbABEL is the only package that implements specific mixed-model based procedures for the study of association in samples with differential relationships, facilitating analysis in family-based studies, studies performed in human genetically isolated populations, and outbred animal populations.

In theory, the mixed model we have described can also be used to correct for population stratification in a study where a number of (population-based and family based) samples come from differentiated genetic populations [12, 19]. However, given the different genetic and potentially different environmental compositions of such differentiated populations, similar heritabilities can not be assumed in all study populations. We speculate that, in practice, one should combine population-specific (fixed or mixed-model) approaches with structured association or similar methods. For example, one could identify sets of individuals coming from divergent genetic populations using either prior information or analysis of the principal components of the genomic kinship matrix [11]; perform standard analysis in population-based sets and mixed-model analysis in family based sets (or those exhibiting substantial cryptic relatedness), as described here; and finally combine the results using meta-analysis. The best strategy to analyze such complex studies is to be addressed elsewhere in more details.

The two-step mixed model-based score test implemented in ProbABEL is an extension of the family-based association score test suggested by Chen and Abecasis [14], and is similar in its logic to the GRAMMAR and GRAMMAR-GC tests described by Aulchenko et al. [3, 15]. In the test procedure, the model is split into two parts (see the equation 2.6 in ”Two-step score test approximation to the mixed model”), the first of which contains the effects of nuisance parameters, including random genetic effects, and the second includes the parameters of interest (SNP effects and SNP-interacting covariates). Estimation in the second step is performed based on the estimates obtained from fitting the first part.

Strictly speaking, the test defined in this manner is correct if the distributions of covariates in the first and the second parts of the model are independent conditional on the estimated phenotypic variance-covariance matrix. This assumption is most likely to be true when the covariates included in the base model are environmental ones, and thus are not expected to exhibit conditional correlation with SNPs. However, when endogenous risk factors, such as body mass index, are included as the covariates in the base model, some SNPs are expected to exhibit covariance with this covariate. In such situations, the covariate should be included in the second step analysis. This, however, may violate the assumptions of the score test if the covariate explains a large proportion of trait variance. In such situation we expect that the test will become conservative and may be less powerful compared to the classical maximum likelihood analysis.

At present, GWA analysis of millions of imputed SNPs using the `-mmscore` option in `ProbABEL` takes a few days for samples of a few thousands of people. However, the relationship between CPU time and the number of subjects is not linear; as the number of subjects reaches 5,000 or more, the mixed-model based analysis will take too much time (weeks to months) when using a single CPU. A straightforward approach to solve this problem would be to use parallel computations. Still, the non-linear dependency of computational time on the number of subjects may become a major analysis bottleneck with larger and larger studies becoming available.

Other software packages which implement similar mixed-model functionality and are suitable for GWA analyses are MERLIN [23] and QxPak [24]. In particular, MERLIN implements the two-step score test [14], which is equivalent to our test in the absence of covariates. QxPak is a flexible tool for mixed modeling of quantitative traits, which implements classical full Maximum Likelihood and Restricted Maximum Likelihood estimation procedures. Neither MERLIN nor QxPak, however, allow for analyses of imputed data in the form of regression onto estimated genotype probabilities. Both packages assume that pedigree structure is known, and estimate kinship based on that.

On the contrary, the input required by ProbABEL consists of the inverse matrix of estimated variances and covariances between the phenotypes of study individuals. This matrix can be obtained in a number of different ways; our standard approach is to estimate it using `GenABEL`'s `polygenic()` function based on kinship estimated from genomic data, as computed with the `ibs(..., weight="freq")` function. However, it is possible and straightforward to use kinship estimated from pedigree data as well (using, e.g., "`kinship`" library of `R`) in the `polygenic()` procedure. The latter approach is preferable in a study where no genome-wide data is available for estimation of genomic kinship (such as a candidate gene or region study).

Presently, there is no package (including `ProbABEL`), which allows for genome-wide association analysis of binary traits or time-till-event outcomes under a mixed model or an approximation to a mixed model accounting for relatedness, and providing the correct estimates of Odds or Hazards Ratios. With the growing number of GWA scans performed in families and genetically isolated populations, this gap needs to be filled.

For population-based analyses using fixed effects models, `ProbABEL` computes Maximum Likelihood estimates of the parameters and the standard errors under the alternative hypothesis, allowing a Wald test for every parameter under consideration. The global SNP significance test is implemented using the Likelihood Ratio Test. Theoretically, the Wald test can be used for the same purpose, thereby avoiding the need to re-estimate the null model with respect to each SNP. However, in GWAS with imputed data, where full information is available for all SNPs, the null model estimation needs to be performed only once, and can then used for testing all SNPs. Thus the overhead related to re-estimation of the null model is minimal, and, for that reason, we did not implement the global SNP significance Wald test.

We should emphasise that, in general, the `ProbABEL` software can be used to do massive regression analyzes using any type of real-type outcomes and predictors. As such, `ProbABEL` is not restricted to SNP, or even, more generally, to genetic analyzes and can be used for any analyzes requiring regression of a dependent variable on a very large number of independent variables in turn. For example, `ProbABEL` may be use to perform association testing among traits and Copy Number Polymorphisms [25].

The practical applicability of `ProbABEL` for the analysis of GWAS is confirmed by the fact that the early versions of the package were successfully used for analysis of multiple data sets, including already published genome-wide analyzes of such various traits as height [26, 27], gout [28], waist circumference [29], smoking initiation [30], and others.

Conclusions

We developed the `ProbABEL` software package, which facilitates fast genome-wide association analysis of imputed data under linear, logistic and Cox proportional hazards models. For quantitative traits, the package also implements a two-step mixed model-based score test for association in samples with differential relationship, facilitating analysis in family-based studies, studies performed in human genetically isolated populations, and outbred animal populations.

Availability and requirements

PROJECT NAME: ProbABEL

PROJECT HOME PAGE: `http://mga.bionet.nsc.ru/~yurii/ABEL/` (source code and binaries for various platforms), http:r-forge.r-project.orgprojectsgenabel (project development page)

OPERATING SYSTEM(S): source code was successfully compiled and used on Windows, Mac OS X, Linux, SUN Solaris

PROGRAMMING LANGUAGE: C, C++, Perl

OTHER REQUIREMENTS: make

LICENSE: GNU GPL

ANY RESTRICTIONS TO USE BY NON-ACADEMICS: None

Authors contributions

YSA developed the original idea, methodology, and code for the fixed effects part. MVS contributed the code for the interaction testing and two-step mixed-model based procedures. CvD provided ERF study data. All co-authors contributed to writing of the manuscript.

Acknowledgements

We thank Dr. Aaron Isaacs, Prof. Tatiana I. Axenovich, Dr. William Astle, Prof. David J. Balding and two anonymous reviewers for valuable suggestions and discussion. We also appreciate the efforts of Thomas Lumley and other member of the R community to provide free statistical computing code. Above all we appreciate help of ProbABEL and GenABEL users who provided valuable comments about the functionality and helped detecting errors in the ProbABEL code.

The work was supported by the Centre for Medical Systems Biology (CMSB), the Netherlands Organization for Scientific Research (NWO) and Russian Foundation for Basic Research (NWO-RFBR 047.017.043) and the Horizon program of the Netherlands Genomics Initiative (93518029).

Tables

Table 2.1: Mean values of the test statistics (Wald for Linear, score for `mmscore`), genomic control λ (median test statistic over 0.455), and type 1 error at different α for different models.

			α		
Model	Mean(T^2)	λ	0.05	0.01	0.001
Linear	1.206	1.224	0.073	0.018	0.0027
Linear, robust	1.210	1.228	0.073	0.018	0.0028
Linear, mmscore	0.984	1.007	0.047	0.009	0.0011

Tests were performed using a trait dependent on two covariates and with (adjusted) heritability of 30%. Only SNPs with estimated minor allele frequency greater than 0.01 ($n = 212,691$) used. Linear: standard linear model; Linear, robust: linear models using with standard errors; Linear, mmscore: two-step approximation to mixed model, fixed effects included in step 1 of analysis.

Table 2.2: Time for analysis of chromosome 2 imputed data (220, 833 SNPs)

Model	Option	No. people	CPU time
Linear	–	500	0m 43s
		1000	1m 23s
		1500	2m 10s
Linear	`-robust`	500	0m 50s
		1000	1m 43s
		1500	2m 35s
Linear	`-mmscore`	500	16m 18s
		1000	92m 45s
		1500	231m 49s
Logistic	–	500	3m 20s
		1000	6m 38s
		1500	10m 8s
Logistic	`-robust`	500	3m 25s
		1000	6m 53s
		1500	10m 29s
Cox PH	–	500	2m 18s
		1000	4m 30s
		1500	6m 43s

In all analyzes, 2 covariates were included in the model

BIBLIOGRAPHY

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106,** 9362–9367 (June 2009).
2. 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).
3. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).
4. 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).
5. Clayton, D. & Leung, H.-T. An R package for analysis of whole-genome association studies. *Hum Hered* **64,** 45–51 (2007).
6. Li, Y & Abecasis, G. Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. *American Journal of Human Genetics* **S79,** 2290 (2006).
7. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10,** 387–406 (2009).
8. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3,** e114 (2007).
9. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5,** e1000529 (2009).
10. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
11. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38,** 904–909 (2006).
12. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38,** 203–208 (2006).
13. Aulchenko, Y. S., de Koning, D.-J. & Haley, C. Genomewide rapid association using mixed model and regression: afast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177,** 577–585 (2007).
14. Chen, W.-M. & Abecasis, G. R. Family-based association tests for genomewide association scans. *Am J Hum Genet* **81,** 913–926 (2007).

15. Amin, N., van Duijn, C. M. & Aulchenko, Y. S. A genomic background based method for association analysis in related individuals. *PLoS One* **2,** e1274 (2007).

16. McCulloch, C. E. & Searle, S. R. *Generalized, Linear, and Mixed Models* (John Wiley & Sons, Inc, 2001).

17. White, H. A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica* **48,** 817–838 (1980).

18. Zeileis, A. Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software* **11,** 1–17 (2004).

19. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*pages.

20. Pardo, L. M., MacKay, I., Oostra, B., van Duijn, C. M. & Aulchenko, Y. S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69,** 288–295 (2005).

21. Anderson, C. A. *et al.* Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am J Hum Genet* **83,** 112–119 (2008).

22. Hao, K., Chudin, E., McElwee, J. & Schadt, E. E. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet* **10,** 27 (2009).

23. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30,** 97–101 (2002).

24. Perez-Enciso, M. & Misztal, I. Qxpak: a versatile mixed model application for genetical genomics and QTL analyses. *Bioinformatics* **20,** 2792–2798 (2004).

25. Kim, W., Gordon, D., Sebat, J., Ye, K. Q. & Finch, S. J. Computing power and sample size for case-control association studies with copy number polymorphism: application of mixture-based likelihood ratio test. *PLoS One* **3,** e3475 (2008).

26. Axenovich, T. I. *et al.* Linkage analysis of adult height in a large pedigree from a Dutch genetically isolated population. *Hum Genet* **126,** 457–471 (2009).

27. Estrada, K. *et al.* A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. *Hum Mol Genet* **18,** 3516–3524 (2009).

28. Woodward, O. M. *et al.* Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *Proc Natl Acad Sci U S A* **106,** 10338–10342 (2009).

29. 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).

30. Vink, J. M. *et al.* Genome-wide association study of smoking initiation and current smoking. *Am J Hum Genet* **84,** 367–379 (2009).

2.2 DETECTING LOW FREQUENT LOSS-OF-FUNCTION ALLELES IN GENOME WIDE ASSOCIATION STUDIES WITH RED HAIR COLOR AS EXAMPLE

***Fan Liu*[1]*, Struchalin M*[2]*, Kate van Duijn*[1]*, Albert Hofman*[2]*, Andre G. Uitterlinden*[2,3]*, Cornelia van Duijn*[2]*, Yurii S. Aulchenko*[2]*, Manfred Kayser*[1]**

[1] *Department of Forensic Molecular Biology, Erasmus University Medical Center, Rotterdam, The Netherlands,*
[2] *Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands,*
[3] *Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands.*

Abstract

Multiple loss-of-function (LOF) alleles at the same gene may influence a phenotype not only in the homozygote state when alleles are considered individually, but also in the compound heterozygote (CH) state. Such LOF alleles typically have low frequencies and moderate to large effects. Detecting such variants is of interest to the genetics community, and relevant statistical methods for detecting and quantifying their effects are sorely needed. We present a collapsed double heterozygosity (CDH) test to detect the presence of multiple LOF alleles at a gene. When causal SNPs are available, which may be the case in next generation genome sequencing studies, this CDH test has overwhelmingly higher power than single SNP analysis. When causal SNPs are not directly available such as in current GWA settings, we show the CDH test has higher power than standard single SNP analysis if tagging SNPs are in linkage disequilibrium with the underlying causal SNPs to at least a moderate degree ($r^2 > 0.1$). The test is implemented for genome-wide analysis in the publically available software package GenABEL which is based on a sliding window approach. We provide the proof of principle by conducting a genome-wide CDH analysis of red hair color, a trait known to be influenced by multiple loss-of-function alleles, in a total of 7,732 Dutch individuals with hair color ascertained. The association signals at the *MC1R* gene locus from CDH were uniformly more significant than traditional GWA analyses (the most significant P for CDH $= 3.11 \cdot 10^{-142}$ vs. P for rs258322 $= 1.33 \cdot 10^{-66}$). The CDH test will contribute towards finding rare LOF variants in GWAS and sequencing studies.

PLoS One,
2011 Nov 29, 6(11)

Genome-wide association studies (GWAS) have successfully identified thousands of common variants associated with many complex human phenotypes including common diseases (www.genome.gov/gwastudies). However, with a few exceptions, common variants identified to date explain only a small fraction of the overall heritability of the traits studied. It was speculated that searching for common variants with increasingly smaller effects are unlikely to substantially account for the missing heritability [1]. Thus, there have been calls for shifting the attention from genome scans of larger samples to studies of rarer variants with larger effect [2]. In particular, it has been proposed that heterozygous loss-of-function (LOF) variants may account for an essential portion of the missing heritability [1, 3].

LOF variants represent alleles resulting in reduced or abolished protein function by disrupting not only the protein-coding genes but also any essential genetic element, including non-coding regulatory motifs. They have a variety of forms, including singlebase substitutions such as nonsense SNPs or splice site disruptions and small or larger insertions/deletions that change the reading frame or remove an entire gene. These are surprisingly common in healthy individuals in that gene disruption, as a result of positive selection, can be beneficial [4, 5]. On the other hand, people at the extremes of trait distributions are more likely to carry trait-associated LOF variants [6]. LOF variants are mostly recognized by their genetic association with a variety of phenotypes largely inherited in a recessive manner. It is important to note multiple LOF variants at the same locus can act not only in the homozygote state, but also in the compound heterozygote (CH) state, where the presence of two different LOF variant alleles at the same gene, one on each homologue chromosome, influence the phenotype. In such cases, the CH state would be much more frequent than the homozygote state for any individual variant. We thus expect a power gain by taking the CH state into account in GWAS or in genome sequencing studies.

There are numerous convincing examples that multiple LOF variants in a gene collectively influence a phenotype. Some examples are *HFE* and hemochromatosis [7], *PLA2G7* and coronary heart diseases [8], *SLC22A12/SLC2A9* and renal hypouricemia [9, 10], *KCNQ1* and Jervell and Lange-Nielsen syndrome [11], *NCCT* and Gitelmans syndrome [12], *ABCC6/GGCX* and pseudoxanthoma elasticum [13, 14, 15], TG and congenital goiter [16], *SCN5A* and Brugada syndrome [17], *P2RX7* and inflammatory response [18], *ABCA12* and congenital ichthyoses [19], *TRIM32* and nephrogenic diabetes insipidus [20], *WFS1* and Wolfram syndrome [21], and *CLDN16* and hypomagnesaemia [22]. Through this study we use LOF variants in *MC1R* and red hair color as an example where empirical data were available. Polymorphisms leading to complete loss of function of *MC1R* are responsible for the red hair/fair skin pigmentation phenotype [23], characterized by tendency to burn and inability to tan, and has been significantly linked to the development of UVinduced skin cancer, in particular melanoma. At least 9 distinct variants in *MC1R* contribute to an increased chance of developing red hair [23, 24, 25, 26]. The relative chance for the red hair phenotype was estimated to be, in general, 15-fold greater among the individuals carrying any single variant allele, compared to noncarriers, and 170-fold higher among homozygotes or CH carriers [23]. A GWAS in $2,986$ Icelanders based on the Illumina 317K chip successfully confirmed the association between red hair and variants in *MC1R* where the most significant signal was derived from a tagging SNP (rs4785763 $P = 3.2 \cdot 10^{-56}$) [26]. The authors subsequently achieved a much stronger association by additionally genotyping two nonsynonymous SNPs not assayed on this chip (rs1805007 $P = 2.0 \cdot 10^{-142}$, rs1805008 $P = 4.2 \cdot 10^{-95}$). The fact that the causal alleles have an extraordinarily large effect which is sufficiently frequent in European populations (0.142 for rs1805007 and 0.108 for rs1805008 in HapMap CEU) allowed successful detection of the genome-wide significant signals from these tagging SNPs. However, in more common situations such LOF variants may have smaller effect sizes and can occur at lower frequencies, and so be undetectable, even if they are directly observed through the next generation sequencing techniques. Relevant statistical methods for detecting and quantifying their effects are sorely needed.

It was speculated that an increased statistical power may be achieved by analyzing multiple neighboring low-frequency variants simultaneously. Several methods have been proposed for analyzing a collection of selected rare mutations to test for group-wise association with a disease status. Recent developments in this area include the cohort allelic sums test (CAST) [27], the combined multivariate and collapsing (CMC) method [28], and the weighted sum statistic (WSS) [29]. In the CAST method, the overall frequency of all exonic alleles in a gene is compared between cases and controls. In the CMC method, all selected rare variants are collapsed and treated as a single common variant allele. The WSS method jointly analyzes a group of rare mutations to test for an excess of mutations in cases. Madsen et al. [29] compared the performance of CAST, CMC, and WSS and showed that WSS was the most powerful under four genetic models. In general, the power of these methods depends on the portion and the frequency of causal variants included. However, none of these methods focused on the CH and they are most suitable for analyzing exonic regions with a collection of rare and possibly functional alleles.

Here, we aim to develop a computationally efficient method to screen for multiple LOF variants, which does not rely on function annotation. The performance of this method is evaluated based on simulated phenotypes and real genotypes from the Illumina 550K chip available for $10,213$ Dutch individuals from the Rotterdam Study, and compared with single SNP analysis and WSS. Finally, we provide a proof of principle using a GWAS of red hair in $7,732$ participants who provided information on their hair color.

Rotterdam Study, microarray genotypes, and hair color data

The Rotterdam Study (RS1) [30] has been in operation since 1990 and initially included 7,983 participants living in Rotterdam, The Netherlands. The RS2 [31] is an extension of the cohort, started in 1999 and includes 3,011 participants. The RS3 [32] is a further extension of the cohort started in 2006 and includes 3,932 participants. RS1 and RS2 were genotyped using the Infinium II HumanHap550 K Genotyping BeadChip version 3 and RS3 was genotyped using Human 610 Quad Arrays of Illumina. Collection and purification of DNA, genotyping, imputation, merging, and quality control details have been described before [33, 34]. Hair color was collected in RS1 and RS2 by means of a questionnaire, with self reporting of 4 hair color categories; fair, brown, red, or black when young. After quality control, this study included a total of 10,213 individuals with 550 K genotyped SNPs, among whom 7,732 individuals provided hair color information (*N* red hair = 241). The Medical Ethics Committee of Erasmus Medical Center, Rotterdam, approved this study. All participants provided written informed consent.

MC1R *SNP genotyping*

Multiple LOF mutations in *MC1R* cause red hair color. These mutations are largely recessive when considered individually and interact with each other in compound heterozygotes. Two SNPs rs1805007 (R151C) and rs1805008 (R160W) known to have the largest effects [24] but not present on the Illumina 550 K chip, were genotyped separately using melt curve genotyping. The assay design and primer synthesis were done by Tib Molbiol (Berlin, Germany, Table S1). For laboratory details see Text S1.

Expected P values from the CDH test of 2 causal SNPs

The expected P values from the CDH test of 2 causal SNPs was mathematically derived as described below (also illustrated using an excel macro Table S2). Consider two physically close SNPs with low MAFs (1.5%). When their LD is low (as measured by r^2), they approximately and independently follow HWE. The frequency of the combined genotypes is expected to follow:

$$R = \left\{ \begin{array}{ccc} (1-q_1)^2(1-q_2)^2 & 2(1-q_1)^2(1-q_2)q_2 & (1-q_1)^2 q_2 \\ 2(1-q_1)q_1(1-q_2)^2 & 4(1-q_1)q_1(1-q_2)q_2 & 2(1-q_1)q_1 q_2^2 \\ q_1^2(1-q_2)^2 & 2q_1^2(1-q_2)q_2 & q_1^2 q_2^2 \end{array} \right\}$$

where q_1 and q_2 are the frequencies of minor alleles. Note here because q_1 and q_2 are small, $R_{(2,3)}$, $R_{(3,2)}$, and $R_{(3,3)}$ are close to zero. The CH state $R_{(2,2)}$ is more frequent than the homozygote state of either SNP ($R_{(1,3)}$ and $R_{(3,1)}$), for example, when $q_1 = q_2$,

$$R_{(2,2)}/R_{(1,3)} = R_{(2,2)}/R_{(3,1)} = 4.$$

Consider a genetic model in which the homozygote and compound heterozygote genotypes lead to an increased prevalence of a binary phenotype, so that the joint penetrance table of the two SNPs can be modeled using a baseline prevalence α, together with a GRR, denoted as β here.

$$F = \begin{Bmatrix} \alpha & \alpha & \alpha\beta \\ \alpha & \alpha\beta & \alpha\beta \\ \alpha\beta & \alpha\beta & \alpha\beta \end{Bmatrix}$$

Given the total sample size n, the expected genotype count in cases is the element by element multiplication of R and F

$$D = nRF,$$

as well as in controls

$$U = nR(1 - F).$$

Note here we consider population based studies typically consist of a large number of healthy individuals and a small number of cases in terms of rare diseases or extreme phenotypes. This is different from the conventional case-control designs where subjects are selected based on the status of a particular disease. Therefore, n needs to be sufficiently large to reach reasonable power, for example, one would need 10,000 population samples to obtain 500 cases for a phenotype with 5% prevalence. However, the definitions of D and U can be easily modified if the number of cases and controls are fixed by design.

Based on F, a two-by-two contingency table can be formed by collapsing the lower triangle cells in both cases and controls

$$O = \begin{Bmatrix} \sum D - (D_{(1,1)} + D_{(1,2)} + D_{(2,1)}) & \sum U - (U_{(1,1)} + U_{(1,2)} + U_{(2,1)}) \\ D_{(1,1)} + D_{(1,2)} + D_{(2,1} & U_{(1,1)} + U_{(1,2)} + U_{(2,1)} \end{Bmatrix}$$

In single SNP analysis, a two-by-three table can be formed,

$$O = \begin{Bmatrix} \sum_{i=1}^{3} D_{1,i} & \sum_{i=1}^{3} U_{1,i} \\ \sum_{i=1}^{3} D_{2,i} & \sum_{i=1}^{3} U_{2,i} \\ \sum_{i=1}^{3} D_{3,i} & \sum_{i=1}^{3} U_{3,i} \end{Bmatrix}$$

where O is an expected matrix of counts under the alternative hypothesis (not confused with real observations). The Chi-square value is computed using standard operations for contingency tables,

$$c = \sum \frac{(O - E)^2}{E}$$

which follows the Chi-square distribution with 1 df for CDH test and 2 df for a single SNP test. The expected P values from the CDH analysis of causal SNPs are compared with those from the single SNP analysis under comparable parameters, in which we set $q = q_1 = q_2$ for illustration purposes.

SNP sampling and trait simulation

Two physically close (< 200 kb) SNPs $S1$ with alleles a and A (frequency of A 1% to 5%) and $S2$ with alleles b and B (frequency of B 1% to 5%) were randomly sampled $10,000$ times without replacement over the Illumina 550 K chip in the Rotterdam Study (N individuals = $10,213$). The r^2 values between SNPs a and b are derived (Supplementary Figure 4.1). For each SNP pair, we simulated a set of binary trait status at the fixed baseline prevalence of 5% under various GRR ranging from 1 to 10, where $GRR_{AA} = GRR_{BB} = GRR_{aAbB}$. The $GRR = 1$ represents the null hypothesis of no genetic association. The tagging SNP $S3$ with alleles c and C is selected if it is in LD with $S1$ and the tagging SNP $S4$ with alleles d and D is selected if it is in LD with $S2$ based on various r^2 thresholds (ranging from 0 to 1) without any constraint on MAF. The SNPs $S1$, $S2$, $S3$, and $S4$ were tested for association with the simulated trait separately using a Chi-squared test with 2 df. The CDH test was conducted for the collapsed genotypes between SNPs $S1$ and $S2$ and between $S3$ and $S4$ using Chi-squared test with 1 df.

Compare CDH and WSS

Madsen et al. [29] have compared the performance of the CAST, CMC, and WSS methods for testing associations involving rare variants and showed that WSS was the most powerful under four genetic models: recessive-set, recessive, additive and dominant. The recessive-set model is the same model as considered in this study. We compared the power of CDH with WSS under recessive-set model using simulations and focus on the scenarios whether or not causal SNPs were directly observed. Under both scenarios, the proportion of causal SNPs in a genomic region is variable and other parameters are fixed ($GRR = 10$, N individuals $= 10,000$, $a = 0.05$). WSS was implemented as described in [29] using a permutation correction of $k = 1000$ as suggested. The CDH test was conducted in a pair-wise manner and the minimal P value was Bonferroni corrected by the total number of tests ($n(n21)/2$). The P-value threshold of 0.05 was used for rejecting the null hypothesis of no association. A region spanning 200 kb was randomly sampled $10,000$ times over the Illumina 550 K chip. For each sampling, a binary trait was simulated by considering a portion of the low frequency variants ($MAF < 0.05$) in the region to be causal under the recessive-set model. Other parameters were fixed ($\alpha = 0.05$, $N = 10,000$, and $GRR = 10$ for carriers of any homozygote or CH genotype of the causal variants). Four scenarios were investigated where (1) all SNPs in the region were analyzed by CDH, (2) all SNPs with $MAF < 0.05$ were analyzed by WSS, (3) all non-causal SNPs were analyzed by CDH, and (4) all non-causal variants with $MAF < 0.05$ were analyzed by WSS.

We also compared WSS with CDH using empirical hair color data. The *MC1R* region from 87.88 to 88.69 Mb on chromosome 16 encompassed 90 genotyped SNPs with call rate $> 95\%$ and was selected for testing association with red hair using CDH and WSS. The two additionally genotyped causal SNPs rs1805007 and rs1805008 were included in or excluded from the region, mimicking the scenarios where causal variants are directly available or not. SNPs in this region were cumulatively included into the WSS analysis according to their MAF in ascending order. All SNPs in the *MC1R* region were analyzed by CDH in a pair-wise manner, and the minimal P value was Bonferroni corrected by the total number of tests ($n = 4005$).

GWA analysis

The GWA analysis was conducted using GenABEL [35] and followed closely the methods previously described [34]. The inflation factor for red hair color was 1.01. Adjusting for gender and the main principal components from the multidimensional scaling analysis did not alter GWA results. Age was adjusted at the stage of phenotype ascertainment (recalled hair color when young). Single SNP analysis was performed using a score test (*qtscore*) in GenABEL with $2df$. In order to check if CH, rather than double heterozyotes, may indeed explain the identified association, we inferred haplotypes using the expectation maximization algorithm implemented in R library *haplo.stats* [36]. All SNPs in this study were annotated according to the NCBI genome-build version 36.3.

Results

The CH model

We consider a genetic model mimicking the situation where recessive and CH genotypes of two low-frequent variants are responsible for the genetic association with a binary phenotype (Figure 2.1). Consider two SNPs with common alleles a and b and minor alleles A and B, which are causal and of low frequency ($1\% < MAF < 5\%$). Each SNP is largely recessive when considered individually, meaning homozygotes for any of the causal alleles (AA or BB) leads to an increased genotypic relative risk (GRR) of expressed phenotype (Figure 2.1). When two SNPs are considered jointly, not only the homozygote genotypes but also the CH genotype ($AaBb$) leads to an increased GRR. Here the causal alleles A and B are assumed to reside on different haplotypes as suggested previously [3, 4, 10, 11], meaning frequencies of the $AABb$, $AaBB$, and $AABB$ genotypes are close to zero. We examined this assumption empirically using the r^2 value, because a low r^2 value would indicate A and B resided on different haplotypes. Note D', another frequently used LD measurement, is not necessarily low or high when A and B alleles reside on different haplotypes. A pair of two physically close (< 200 kb) and low-frequent ($1\% < MAFs < 5\%$) SNPs was resampled (N resampling = $10,000$) over the genome (Illumina 550 K chip) in the Rotterdam Study (N individuals = $10,213$). The majority (56.3%) of SNP pairs showed very low r^2 (< 0.01, Supplementary Figure 4.1). For the SNP pairs with low r^2, the joint genotypes $aaBB$ (on average 0.11%), $AAbb$ (0.11%), and $AaBb$ (0.25%) were small and the frequencies of $AABb$, $AaBB$, and $AABB$ were close to zero (Figure 2.1). The frequency of the $AaBb$ genotype was on average 2.27 times higher than that of the $AAbb$ or $aaBB$ genotypes. About 18% SNP pairs were in high LD ($r^2 < 0.9$, Supplementary Figure 4.1). Because the cross genotypes for two SNPs in high LD provided little or no additional information than that provided by either SNP alone (Supplementary Figure 4.2), our model focuses on the low r^2 scenario. We developed a simple test, named the collapsed double heterozygote (CDH) test, to detect the association caused by this particular genetic model.

CDH test of two causal SNPs

We first considered the scenario where two causal SNPs are directly genotyped, which is notably unrealistic for SNP microarray data but may be the case in next generation genome sequencing studies. The CDH test is based on the Chi-squared test as defined below. We denote the two causal SNPs as $S1$ with alleles a and A and $S2$ with alleles b and B. Let D and U be observed genotype counts in cases and controls and both follow a 3-by-3 matrix form,

$$\left\{\begin{array}{rcc} aabb & aAbb & AAbb \\ aabB & aAbB & AAbB \\ aaBBaABBAABB & & \end{array}\right\}$$

The observed matrix of counts is collapsed as,

$$O = \left\{\begin{array}{cc} \sum D - (D_{(1,1)} + D_{(1,2)} + D_{(2,1)} & \sum U - (U_{(1,1)} + U_{(1,2)} + U_{(2,1)}) \\ D_{(1,1)} + D_{(1,2)} + D_{(2,1} & U_{(1,1)} + U_{(1,2)} + U_{(2,1)} \end{array}\right\}$$

The Chi-square value is computed using standard operations for contingency tables,

$$c = \sum \frac{(O-E)^2}{E}$$

which follows the Chi-square distribution with 1 df. Note that there is an essential difference in the way that the genotypes are collapsed when tagging SNPs are analyzed (see the subsection of tagging SNPs).

The expected P values from the CDH analysis of two causal SNPs and from the single SNP analysis were mathematically derived as a function of total sample size N, minor allele frequencies of causal SNPs q ($q1 = q2$ for simplicity), and

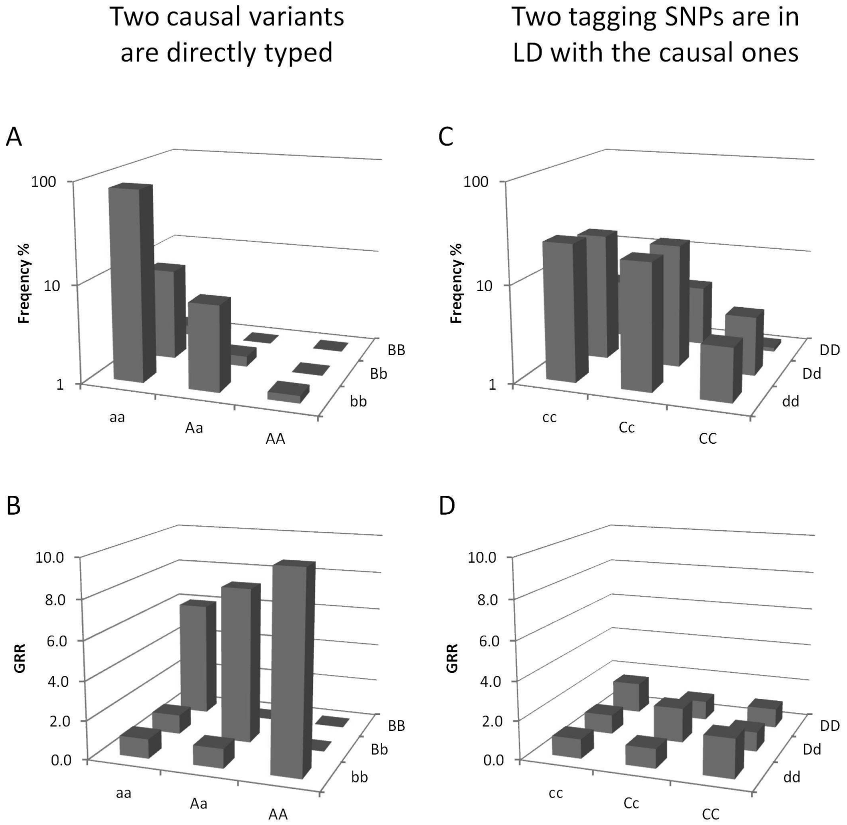
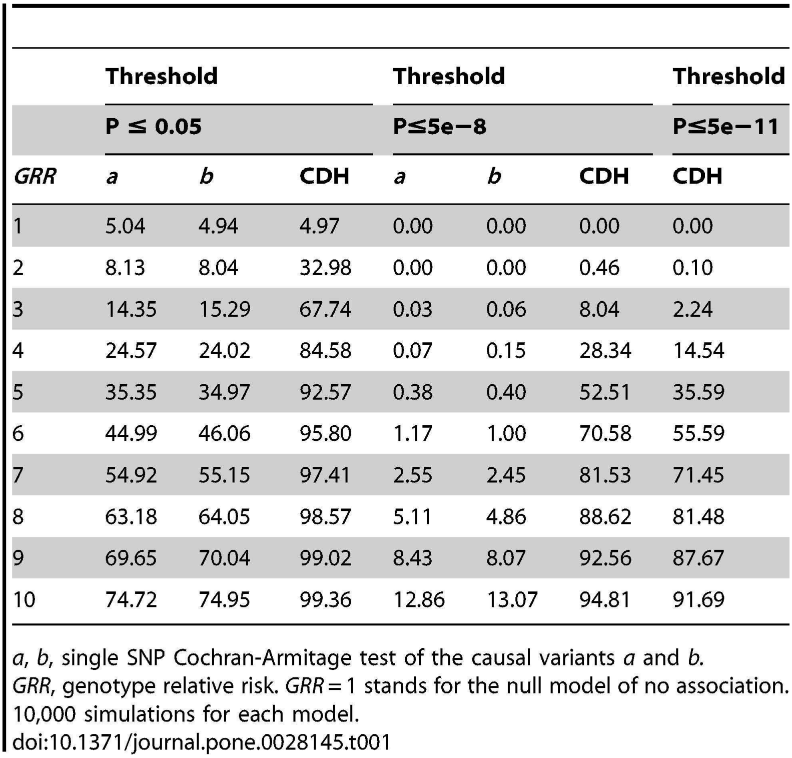


Figure 2.1: **A recessive and compound heterozygote model of the phenotype.** At left part of the figure (A and B) two rare recessive variants at the same gene locus are assumed to be directly genotyped. At the right part of the figure (C and D) two non-causal SNPs with higher minor allele frequencies and in LD with the causal SNPs are genotyped. The upper part of the figure depicts the logarithm scaled frequency of the cross genotypes of two variants (A and C). The lower part of the figure is an example of the genetic model under illustrative parameters. $GRR_{AA} = 8$, $GRR_{AaBb} = 7$, $GRR_{BB} = 6$, $r^2_{ac} = r^2_{bd} = 0.1$ (B and D).

Table 2.3: **Percentage of P values smaller than or equal to the test threshold for single SNP analysis and collapsed genotype analysis of two causal variants.**



	Threshold			Threshold			Threshold
	P ≤ 0.05			P≤5e−8			P≤5e−11
GRR	*a*	*b*	CDH	*a*	*b*	CDH	CDH
1	5.04	4.94	4.97	0.00	0.00	0.00	0.00
2	8.13	8.04	32.98	0.00	0.00	0.46	0.10
3	14.35	15.29	67.74	0.03	0.06	8.04	2.24
4	24.57	24.02	84.58	0.07	0.15	28.34	14.54
5	35.35	34.97	92.57	0.38	0.40	52.51	35.59
6	44.99	46.06	95.80	1.17	1.00	70.58	55.59
7	54.92	55.15	97.41	2.55	2.45	81.53	71.45
8	63.18	64.05	98.57	5.11	4.86	88.62	81.48
9	69.65	70.04	99.02	8.43	8.07	92.56	87.67
10	74.72	74.95	99.36	12.86	13.07	94.81	91.69

a, *b*, single SNP Cochran-Armitage test of the causal variants *a* and *b*.
GRR, genotype relative risk. *GRR* = 1 stands for the null model of no association.
10,000 simulations for each model.
doi:10.1371/journal.pone.0028145.t001

GRR when the base line prevalence of the phenotype a is fixed at 5%. Under the CH model the CDH analysis would be expected to give more significant P values than single SNP analyses (Figure 2.2). For example, with $N = 10,000$, and $0.02 < q < 0.05$, the CDH analysis is expected to give genome-wide significant P values ($< 10^{-8}$) for detecting reasonably large effect sizes $GRR > 3$. With the same sample size, it requires higher minor frequencies ($q >= 0.05$) and larger effect sizes ($GRR > 5.5$) for the single SNP analyses to become genome-wide significant. Note this CDH test gives less significant P values than single SNP analysis for other genetic models where $AaBb$ has no effect. For example, consider 2 independent recessive SNPs or single SNP effect (Supplementary Figure 4.3). We then evaluated the type-1 error rate and the statistical power for CDH using the real genotypes from the Rotterdam Study and simulated phenotypes (Table 2.3).

The type-1 error rates from CDH and the single SNP analysis, whether under the additive or recessive models, were both consistent with the expected under the null hypothesis of no association ($\approx 5\%$ P values smaller than 0.05). Under the alternative hypothesis ($GRR > 1$), the CDH test showed much higher power than the single SNP analyses. For example, at $GRR = 5$, CDH had 52.5% power whereas single SNP analysis had less than 1% power at the significance threshold of $5 \cdot 10^{-8}$ (Table 2.3). The gain in power using the collapsed genotypes was overwhelming even when a much more stringent threshold of $5 \cdot 10^{-10}$ was applied only for CDH (Table 2.3). This extra adjustment allows additional multiple testing in real applications, such as genome-wide implementations based on a sliding window approach or regional implementations based on a pair-wise testing approach (see implementation subsection).

CDH test of two tagging SNPs

A more realistic scenario in GWAS based on SNP microarrays consisting of mainly common variants is that only non-causal tagging SNPs were available. For this scenario we considered two tagging SNPs, $S3$ with alleles c and C and $S4$ with alleles d and D. The tagger $S3$ was selected if it was in LD with $S1$, and the tagger $S4$ was selected if it was in LD with $S2$ based on various r^2 thresholds without constraints on MAF. For a given SNP with $MAF < 5\%$ on the Illumina 550 K chip, there was a good chance (on average 72.74%) of obtaining at least one SNP with an $r^2 > 0.1$ from its 100 neighboring SNPs. The chance of obtaining at least one SNP with $r^2 > 0.5$ was much lower (on average 25.26%). The joint penetrance table for tagging SNPs showed a distinct interaction pattern differing from those previously considered for unlinked loci [37]. An important empirical finding was that only the off-diagonal cells in the cross-genotype table showed any increased GRR, but the $CCDd$, $CcDD$, and $CCDD$ carriers did not have an increased GRR (Figure 2.1). This feature, which appeared to be an antagonistic interaction, can be explained by the very low frequency of the AB haplotypes (also see the subsection of the hair color analysis). This indicates the CDH test is preferred for analysis of the tagging SNPs but the $CCDd$, $CcDD$, and $CCDD$ genotypes should be collapsed together with wildtypes. Again, let D and U be observed genotype counts in cases and controls,

$$O = \begin{Bmatrix} D_{(1,3)} + D_{(2,2)} + D_{(3,1} & U_{(1,3)} + U_{(2,2)} + U_{(3,1)} \\ \sum D - (D_{(1,3)} + D_{(2,2)} + D_{(3,1)}) & \sum U - (U_{(1,3)} + U_{(2,2)} + U_{(3,1)}) \end{Bmatrix}$$

In practice this form can also be used in causal SNP analysis because $AABB$, $AAbB$, and $aABB$ are negligible. Since it was difficult to mathematically derive the expected P values for the CDH test of tagging SNPs, we evaluated type-1 error and power based on simulations. The type-1 error rate for CDH test was consistent with the expected under the null hypothesis of no association ($< 5\%$ nominal P values smaller than 0.05 and 0% smaller than $5 \cdot 10^{-8}$). Under a fixed effect size of the causal SNPs, the most important parameter for power was the r^2 between the causal and tagging SNPs. The product of r^2_{ac} and r^2_{bd} showed a high correlation with the test statistics of CDH (Figure 2.3). As long as $r^2_{ac} \cdot r^2_{bd} > 0.1$,

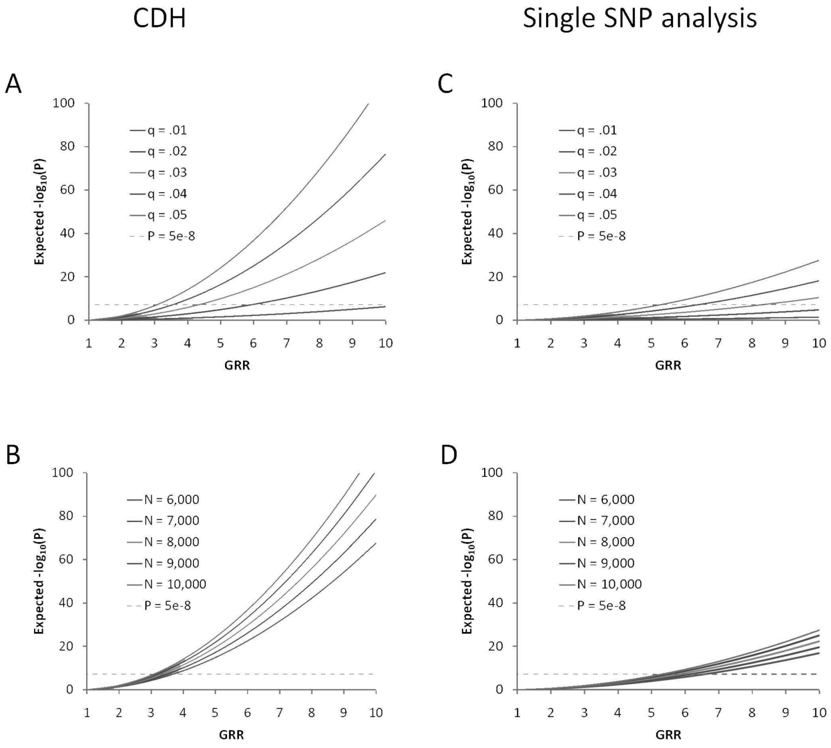


Figure 2.2: **The expected P values for the CDH test.** The $-\log_{10}(P)$ values for two causal SNPs (on the left part of the figure, A and B) and for the single SNP chi-squared test (on the right part, C and D) are derived as a function of the genotype relative risk ($GRR_{AA} = GRR_{BB} = GRR_{AaBb}$ ranging from 1 to 10), the minor allele frequencies ($q = q_1 = q_2$ ranging from 0.01 to 0.05 when N is fixed at $10,000$; A and C), and the total sample size N (ranging from $6,000$ to $10,000$ when q is fixed at 0.05; B and D). The base line prevalence of a binary phenotype is fixed at 5% in all analyses.

the CDH test showed a power considerably higher than single SNP association (Figure 2.3). In particular, when $r^2_{ac} \cdot r^2_{bd} > 0.5$ the CDH had 27% to 91% power to detect a reasonably large effect size ($GRR >= 5$) at the genome-wide significance level ($P < 5 \cdot 10^{-5}$) whereas the single SNP analysis only had poor power (< 10%, Figure 2.3). When $r^2_{ac} \cdot r^2_{bd}$ approached 1, the collapsed tagging SNPs became identical to the collapsed causal SNPs and the power of CDH reached that of the causal SNPs listed in Table 2.3. Finally, a higher power was achieved more often when $MAFs$ of $S3$ and $S4$ were close to that of $S1$ and $S2$ as expected from the relationship between r^2 and $MAFs$. We further compared power of CDH with WSS through simulations. In general, the power of WSS increased when the portion of causal variants included was increased whereas CDH was much less influenced by this parameter and outperformed WSS under all scenarios investigated (Figure 2.4). The most interesting scenario is when the portion of causal variants was low (< 0.1) and the causal variants were not directly observed. Under this scenario the CDH (power 0.41) clearly outperformed WSS (power 0.10).

Software implementation

We implemented the CDH test in the software R package GenABEL [35, 38] and the core computation was implemented using external C/C++ code. The function was based on a sliding window approach and performs the CDH test for every SNP over the genome with the following n SNPs, which can be specified by the user. The n SNPs are not necessarily in or outside of known genes. The minimal P value from each slide is addressed to the first SNP of this slide and Bonferroni corrected for n tests. The Pearson's chisquared or the Fisher's exact test is used depending on the number of individuals in the smallest cell. The total number of tests is $N \cdot n - n(n-1)/2$, where N is the total number of SNPs on the genome, so for a given chip, the computational time is approximately linear to n. For example, with a dual core processor at 2.5 GHz, screening for 500 K SNPs in $10,000$ individuals could be completed in about 7 hours for $n = 100$ and 14 hours for $n = 200$. This implementation is also practically applicable to imputed data sets and screening for 2 million SNPs could be completed in about 28 hours for $n = 100$ and 56 hours for $n = 200$. The effect of window size is relatively small as long as the SNPs cover ≈ 400 kb region. A window consisting of 100 SNPs is on the safe side for screening chips with $500 - 600$ K SNPs.

A GWAS of red hair

We used the red hair color phenotype as the proof of principle to verify the concept that the use of collapsed genotypes is more capable of detecting the presence of multiple recessive variants at the same gene locus than traditional GWA analysis. A genomewide CDH analysis on red-hair color was conducted in 7732 participants (N red hair $= 241$) of the Rotterdam Study using a window

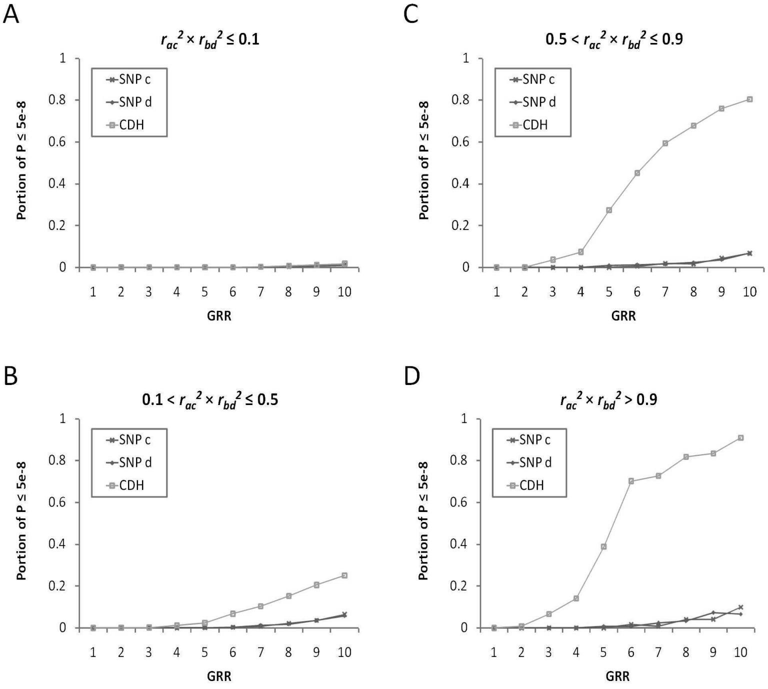


Figure 2.3: **The power of CDH and single SNP analysis.** Proportion of P values $\leq 5 \cdot 10^{-8}$ from the CDH analysis (green dots) and the single SNP Cochran-Armitage test of two tagging SNPs c (red dots) and d (blue dots). Four SNPs were re-sampled 10,000 times from the Illumina 550 K chip. SNPs a and b were physically close ($<$ 200 kb) and had low MAFs ($<$ 5%). SNP c was in LD with a and SNP d was in LD with b. The genotypic relative risk was simulated according to the genotypes of a and b under the recessive and compound heterozygote model, where $GRR_{AA} = GRR_{BB} = GRR_{AaBb}$. The base-line prevalence of a binary phenotype was fixed at 5%. A, when $r_{ac}^2 \cdot r_{bd}^2 \leq 0.1$; B, when $0.1 < r_{ac}^2 \cdot r_{bd}^2 \leq 0.5$; C, when $0.5 < r_{ac}^2 \leq r_{bd}^2 \leq 0.9$, and D, when $r_{ac}^2 \cdot r_{bd}^2 > 0.9$.

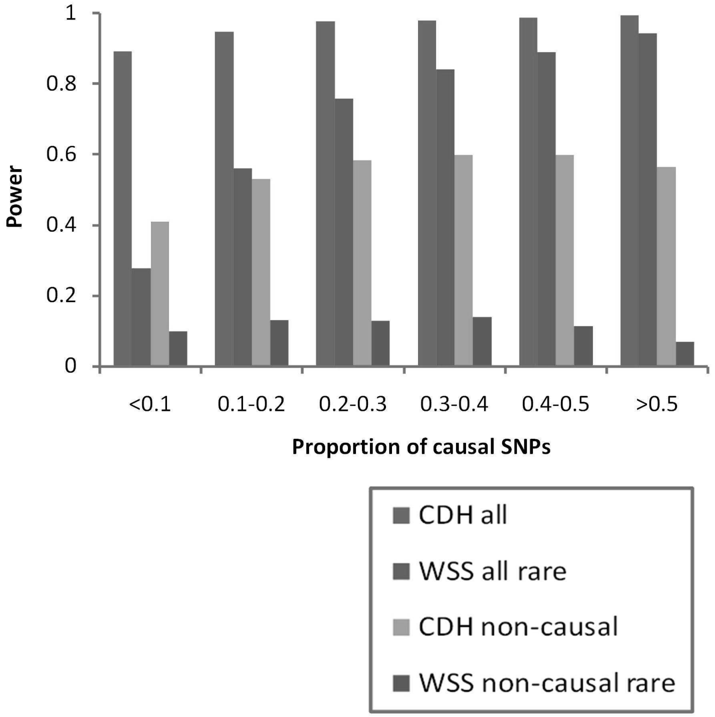
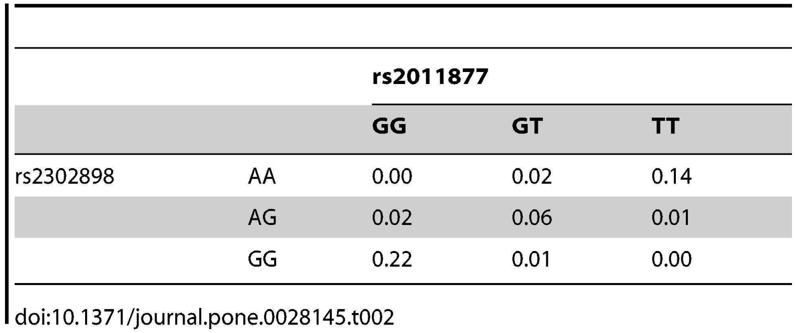


Figure 2.4: **The power of CDH and WSS.** The power of CDH and weighted sum statistic (WSS) [29] was plotted against the portion of causal variants in the sampled region. A region spanning 200 kb was randomly sampled 10, 000 times over the Illumina 550 K chip without replacement. For each sampling, a binary trait was simulated by considering a portion of the rare variants in the region to be causal under the recessive-set model described in [29]. Other parameters were fixed ($\alpha = 0.05$, $n = 10,000$, and $GRR = 10$ for carriers of any homozygote or CH genotype of the causal variants). Four sets of P values were derived when (1) all SNPs in the region were analyzed by CDH (blue), (2) all SNPs with $MAF < 0.05$ were analyzed by WSS (red), (3) all non-causal SNPs were analyzed by CDH (green), and (4) all non-causal variants with $MAF < 0.05$ were analyzed by WSS (purple). The power was defined as the portion of P values smaller than or equal to $5 \cdot 10^{-8}$.

Table 2.4: **Frequency of red hair phenotype as a function of genotype of two non-causal SNPs tagging the causal variants at the *MC1R* gene locus.**



		rs2011877		
		GG	GT	TT
rs2302898	AA	0.00	0.02	0.14
	AG	0.02	0.06	0.01
	GG	0.22	0.01	0.00

doi:10.1371/journal.pone.0028145.t002

size of 100 SNPs (Supplementary Figure 4.4). At chromosome 16, the 87.88 to 88.69 Mb region containing the *MC1R* gene, the association signals from the CDH analyses were uniformly higher than those from single SNP analyses (Figure 2.5).

The most significant P value from CDH after the Bonferroni correction of the window size ($P = 3.11 \cdot 10^{-142}$ between SNPs rs258322 and rs8058895) was markedly more significant than seen with the single SNP association test (P for rs258322 $= 1.33 \cdot 10^{-66}$). On the other hand, there was no inflation of significant results when the hair color phenotype was randomly shuffled 100 times. Besides *MC1R*, no other region showed genome-wide significant evidence where multiple recessive variants were involved (Supplementary Figure 4.4). To further illustrate the underlying mechanism that $CCDd$, $CcDD$, and $CCDD$ carriers did not appear to increase GRR, which might be counterintuitive, we additionally genotyped two important causal SNPs for red hair [24], rs1805007 (R151C) and rs1805008 (R160W), which were not available on the original chip, in the Rotterdam Study population. Figure 2.6 shows diplotypes consisting of these two causal SNPs and two other tagging SNPs for *MC1R* (Figure 2.6). The causal alleles A and B represent rs1805007_T and rs1805008_T, and the tagging alleles C and D for rs2011877_C and rs2302898_T. These two tagging SNPs were selected to not be in very high LD with any causal SNPs for illustration purposes ($r^2_{ab} = 0.007$, $r^2_{ac} = 0.147$, $r^2_{bd} = 0.216$). The $CCDd$ genotype is represented by diplotypes 6 and 13, $CcDD$ by 8 and 14, and $CCDD$ only by 15. This example empirically demonstrated the $A - B$ haplotype at *MC1R* was absent in 7732 individuals. It also explained the unique "antagonistic" interaction expressed in the joint penetrance table of the two tagging SNPs (Table 2.4) where only the offdiagonal cells showed any increased prevalence of red hair.

The CDH test of causal SNPs rs1805007 and rs1805008 resulted in a more significant P value ($P = 4.9 \cdot 10^{-192}$) than testing them separately (P for rs1805007 $= 3.2 \cdot 10^{-139}$, P for rs1805008 $= 3.4 \cdot 10^{-50}$). The CDH test of only tagging SNPs rs2011877 and rs2302898 also resulted in a more significant P value ($P =$

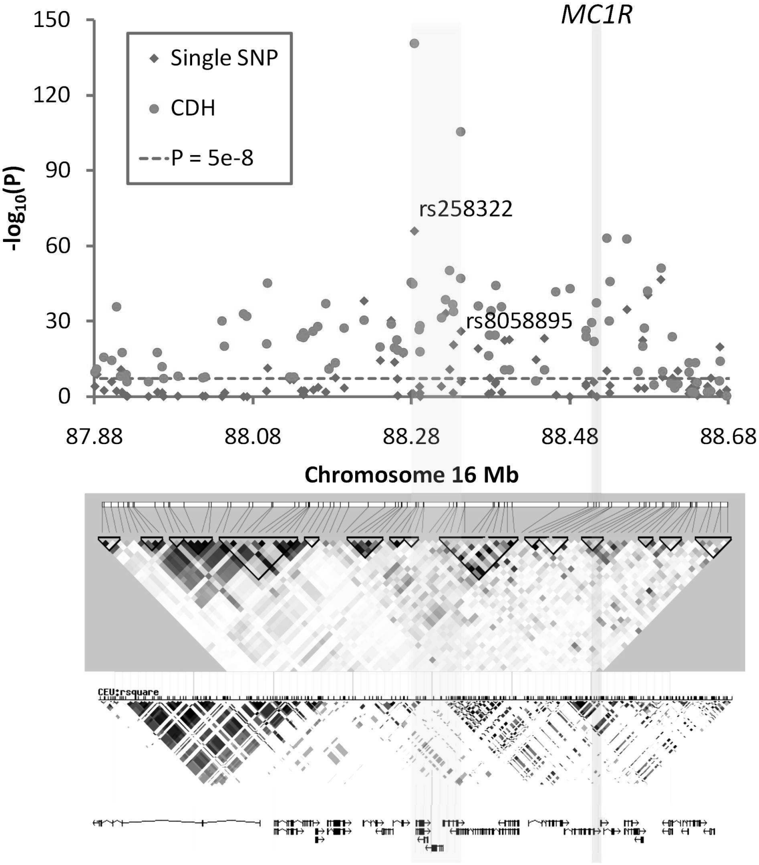
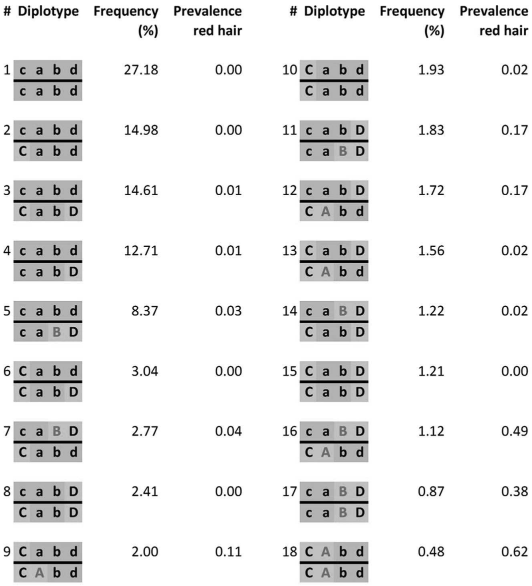


Figure 2.5: **Association between SNPs at *MC1R* and the red hair color in the Rotterdam Study.** The $-\log_{10}(P)$ values for association with red hair color were plotted for each genotyped SNP according to its chromosomal position (blue dots) and for the CDH test in each sliding window consisting of 100 SNPs (green dots represent the left-most SNP). The LD patterns in the Rotterdam Study population and in the HapMap CEU samples (release 27) and the known genes in the region were aligned bellow according to the physical position of the SNPs (genome-build version 36.3). The orange bar indicates the physical position of the *MC1R* gene. The yellow bar indicates the region between two SNPs based on which the most significant P value of the CDH test was obtained (the left-most SNP rs258322 and the right-most SNP rs8058895).



#	Diplotype	Frequency (%)	Prevalence red hair
1	c a b d c a b d	27.18	0.00
2	c a b d C a b d	14.98	0.00
3	c a b d C a b D	14.61	0.01
4	c a b d c a b D	12.71	0.01
5	c a b d c a B D	8.37	0.03
6	C a b d C a b D	3.04	0.00
7	c a B D C a b d	2.77	0.04
8	c a b D C a b D	2.41	0.00
9	C a b d C A b d	2.00	0.11
10	C a b d C a b d	1.93	0.02
11	c a b D c a B D	1.83	0.17
12	c a b D C A b d	1.72	0.17
13	C a b D C A b d	1.56	0.02
14	c a B D C a b D	1.22	0.02
15	C a b D C a b D	1.21	0.00
16	c a B D C A b d	1.12	0.49
17	c a B D c a B D	0.87	0.38
18	C A b d C A b d	0.48	0.62

Figure 2.6: **Frequency of diplotypes and the prevalence of red hair in the Rotterdam Study.** The causal SNP *a* is rs1805007 and *b* is rs1805008. The tagging SNP *c* is rs2011877 and *d* is rs2302898. Causal alleles *A* and *B* are indicated in red color. Common alleles are indicated in green background and minor alleles are indicated in orange background.

$5.9 \cdot 10^{-32}$) than testing them separately (P for rs2011877 $= 6.8 \cdot 10^{-7}$, P for rs2302898 $= 8.9 \cdot 10^{-12}$), confirming a power gain when multiple homozygotes and compound heterozygotes can explain the association. After significant results are obtained for the CDH test of the tagging SNPs, one can further test explicitly that CH genotypes in a collapsed set does have a different effect than the DH genotypes. This test requires diplotype information, which can be inferred statistically. In this example, we compared the red and non-red frequencies in carriers of diplotype 3 against that observed among carriers of diplotypes 7, 12, and 16 (Figure 2.6). The P value derived from this test was also highly significant ($P = 3.9 \cdot 10^{-9}$), pinpointing that CH, but not DH, could account for the identified association. Such analysis can be implemented at the genomewide scale if the whole genome is phased. Finally, diplotypes 9, 11 and 12 seem to have intermediate prevalence compared to the recessive homozygotes or CHs. It is known that multiple causal LOF variants exist in *MC1R* and the two genotyped are the most common of these. Thus, the increased prevalence of diplotypes 9, 11 and 12 can be explained by the CH state of one of these 2 variants with another non-genotyped causal variant in *MC1R*. This also explains that an additive model (Armitage trend test) does not necessarily perform worse than an explicit recessive model in single SNP analysis when more than 2 causal recessive variants exist. Finally, we compared results of CDH with WSS analysis of *MC1R* region (Supplementary Figure 4.5). Using the original chip without causal SNPs rs1805007 and rs1805008, the minimal P value of $1.0 \cdot 10^{-11}$ was obtained for WSS when 7 SNPs with $MAF < 0.07$ were included in the analysis, which is less significant than the P value from the CDH analysis of all SNP pairs in the *MC1R* region (Bonferroni corrected $P = 7.6 \cdot 10^{-142}$). Assuming the two causal SNPs rs1805007 and rs1805008 were available on the chip, the minimal P value ($P = 2.5 \cdot 10^{-19}$) was obtained for WSS when 14 SNPs with $MAF < 0.1$ were included, which was also less significant than the P value obtained from the CDH analysis of all SNPs including the causal ones (Bonferroni corrected $P = 1.6 \cdot 10^{-190}$).

Discussion

We demonstrated theoretically and empirically by simulations that using the collapsed genotypes in GWA analysis is more powerful than single SNP analysis and the WSS method in detecting the presence of multiple LOF variants at a particular gene locus. In a genome scan of the red hair color phenotype this CDH analysis resulted in considerably more significant association signals than single SNP analysis at *MC1R*. Besides *MC1R*, no other region of CH association with red hair was identified. By additional genotyping of two causal SNPs in *MC1R* we confirmed a recessive mechanism underlying this gain in statistical power. The generalizablity of CDH mainly depends on the effect sizes and frequencies of causal alleles. We expect CDH is generalizable to some of the known examples, such as HFE and hemochromatosis, where both the allele effect sizes and frequencies are comparable to *MC1R* alleles. Further, through simulations we showed our method is capable to find LOF alleles with smaller effect sizes ($GRR > 3$) but not with frequencies lower than 1%. It should therefore be emphasized this approach still requires causal alleles to be at some appreciable frequency ($< 1\%$) to be effectively tested and probably not useful for exceptionally rare variants.

Here we focused on a recessive and CH model that addresses, but not restricted to, the SNP interactions caused by LOF variants. This type of SNP interaction is only a subtype of CH-like interactions, e.g. multiple gain function SNPs may well follow the CH model. However, a number of different models exist in theory, in which combinations of different variants influence a particular phenotype. A more "omnibus" hypothesis-testing model may work reasonably well in multiple or most settings. Still, we believe the proposed CH model is valuable. First, it has been suggested that LOF variants are surprisingly common [4, 5] and they may account for a substantial portion of missing heritability [1, 3]. Second, the recessive model is most likely the true model underlying a significant portion of the causal variants undetected by the GWAS conducted to date. In conventional single SNP analysis, the required sample size to detect a recessive allele is a quadratic function of its frequency, which is much larger than the required sample size to detect a dominant or additive allele of the same effect size. This is regardless of the number of causal variants involved at any gene for single SNP analysis. Thus, we expect an essential portion of the currently undetected alleles to be recessive. Third, the magnitude of the power gain of this proposed model is overwhelming for detecting CH-like interactions, in particular for tagging SNP analysis. The more significant P value from the CDH test is clearly driven by the CH carriers. As also shown in the method subsection, when $q_1 = q_2$ the frequency of CH carriers is 4 times higher than homozygote carriers of single SNP, serving as the driving source of the statistical significance. Finally, CDH is computationally simple and practically applicable to large-scale data sets.

It has been repeatedly suggested [28, 29, 39, 40] that rare causal variants are likely to reside on different haplotypes. Under this scenario, the r^2 between two variants is small and the frequency of the AB haplotype is close to zero. Thus, the $AABb$, $AaBB$, and $AABB$ genotype carriers are either unobservable or negligible in practice and the forming of a collapsed marker by collapsing the $AAbb$, $aaBB$, $AaBb$ genotypes has been described in length previously [39, 40]. What has not been so clear is the scenario when tagging SNPs with higher minor allele frequencies are in LD with the rare causal ones, given that the frequency of the CD haplotype is not close to zero. Through simulations and the empirical hair color data we showed that the CD haplotype carriers usually do not have an increased GRR. By grouping the $CcDD$, $CCDd$, and $CCDD$ genotype carriers together with the wild-type carriers, which is the creative element of this paper, we have shown that the tagging SNPs are capable of revealing significant signals.

More importantly, iterative analysis of two tagging SNPs based on a sliding window approach is useful in genome-wide implementations. The proposed models involve only two LOF SNPs in weak LD, but of course one could envision situations in which CH effects could arise due to heterozygosity at a number of different but physically close loci, such as the *MC1R* gene exemplified here or the well-known HLA region. In such cases, iteratively analyzing two of the variants has an advantage over the collection-based methods [27, 28, 29] because power is not compromised by the number of unassociated SNPs included. Although the downside of this method is the additional multiple testing depending on the window size, which must be sufficiently large to cover all SNPs potentially in LD, the power gain is clearly overwhelming. For example, consider the bottom line if the whole genome is tested pair-wise in the genome-scan of red hair color, the CDH test of tagging SNPs would still result in a much more significant P value $(10^{-42} \cdot 10^{12} = 10^{-130})$ than single SNP analysis (10^{-66}) at *MC1R*. On the other hand, for collection-based methods [27, 28, 29], power approaches zero when more and more SNPs are included.

The use of the collapsed genotypes based on tagging or causal SNPs is conceptually distinguished. The interpretation of results may be straightforward when the causal variants are directly available as expected from full genome sequencing data. However, when they are not available and only the tagging SNPs are analyzed, i.e. based on the currently available genotyping chips, the key parameter determining the power is the strength of LD in term of r^2 between the underlying causal SNPs and tagging SNPs. In particular, when $r^2_{ac} \cdot r^2_{bd} > 0.5$ the CDH provides good to excellent power to detect a reasonably large effect size in a population based sample. A critical concern here is the portion of rare variants that are well tagged on the existing genome-wide panels. About 20% of low frequency and physically close SNP pairs from the Illumina 550 k chip have $r^2 > 0.9$ (Supplementary Figure 4.1), and about 25% have $r^2 > 0.5$. These estimates are in line with a recent report showing panels consisting of $300 - 550$ K SNPs capture only a small proportion of the rare nonsynonymous SNPs ($10 - 27\%$ tagged by r^2.0.5) in Europeans [41]. Thus, the portion of rare SNPs tagged by current chips is far from desirable for CDH analysis, except for some candidate traits, such as exemplified here for red hair. Reference panels such as the International HapMap Project [42] (`http://snp.cshl.org/`) and the 1000 Genome Project [43] have already covered up to 7.7 million newly identified rare variants in multiple human populations. The recent progress in the imputation techniques has improved the accuracy of imputing these rare variants [44]. However, in general, the imputation error rate increases as the minor allele frequency decreases across all imputation panels and genotyping chips [45]. On the other hand, using CDH to analyze the denser chips can be safely recommended for screening LOF variants, as in the Illumina 1 M chip, where the density of rare SNPs is already higher than the common ones [39], although full genome sequencing data would be ideal. Finally, regional diplotype analysis is recommended after promising regions are identified with our method. Such promising regions may be followed up by the case selection approach Wang and colleagues have proposed for deep-sequencing [40].

The chi-square statistics used here for analyzing binary traits is simple, and readily extended to general linear models for analyzing quantitative traits with or without covariates. Rather than emphasizing the advances in modern statistics, we underline the known genetic interaction between two or more LOF variants at the same gene: both homozygotes and the CH genotypes result in an increased prevalence of phenotype, and taking this into consideration increases the power in detecting them. The presence of such variants may be common and should be considered in routine analysis in genome scans, particularly for extreme phenotype designs. Our approach is useful in finding these variants in GWAS carried out with chips of ultra-high density, as well as future full genome sequencing studies.

Funding

This study is supported by the Netherlands Genomics Initiative (NGI) / Netherlands Organization for Scientific Research (NWO) (nr. $050-060-810$) within the framework of the Forensic Genomics Consortium Netherlands (FGCN), the Netherlands Forensic Institute and the Research Institute for Diseases in the Elderly (RIDE) (nr. $014-93-015$). The generation and management of GWAS genotype data for the Rotterdam Study ($RS-I$, $RS-II$) was supported by a NWO investments grant (nr. 175.010.2005.011, $911-03-012$). The Rotterdam Study is funded by the Erasmus University Medical Center, the Erasmus University Rotterdam, the Netherlands Organization for the Health Research and Development (ZonMw), the RIDE, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports of the Netherlands, the European Commission (DG XII), and the Municipality of Rotterdam. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Web Resources

GenABEL software for genome-wide analyses: `www.genabel.org`,
A Catalog of Published GWASs: `http://www.genome.gov/gwastudies`.

BIBLIOGRAPHY

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
2. Goldstein, D. B. Common genetic variation and human traits. *The New England journal of medicine* **360,** 1696–1698 (Apr. 2009).
3. Singleton, A. B., Hardy, J., Traynor, B. J. & Houlden, H. Towards a complete resolution of the genetic architecture of disease. *Trends in genetics: TIG* **26,** 438–442 (Oct. 2010).
4. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science (New York, N.Y.)* **312,** 1614–1620 (June 2006).
5. MacArthur, D. G. & Tyler-Smith, C. Loss-of-function variants in the genomes of healthy humans. *Human molecular genetics* **19,** R125–130 (Oct. 2010).
6. Romeo, S. *et al.* Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature genetics* **39,** 513–516 (Apr. 2007).
7. Feder, J. N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature genetics* **13,** 399–408 (Aug. 1996).
8. Song, K *et al.* Sequencing of Lp-PLA2-encoding PLA2G7 gene in 2000 Europeans reveals several rare loss-of-function mutations. *The pharmacogenomics journal,* 425—431 (May 2011).
9. Enomoto, A. *et al.* Molecular identification of a renal urate anion exchanger that regulates blood urate levels. *Nature* **417,** 447–452 (May 2002).
10. Matsuo, H. *et al.* Mutations in glucose transporter 9 gene SLC2A9 cause renal hypouricemia. *American journal of human genetics* **83,** 744–751 (Dec. 2008).
11. Wang, R. *et al.* Novel compound heterozygous mutations T2C and 1149insT in the KCNQ1 gene cause Jervell and Lange-Nielsen syndrome. *International journal of molecular medicine* **28,** 41–46 (July 2011).
12. Gitelman, H. J., Graham, J. B. & Welt, L. G. A new familial disorder characterized by hypokalemia and hypomagnesemia. *Transactions of the Association of American Physicians* **79,** 221–235 (1966).
13. Bergwitz, C. *et al.* SLC34A3 mutations in patients with hereditary hypophosphatemic rickets with hypercalciuria predict a key role for the sodium-phosphate cotransporter NaPi-IIc in maintaining phosphate homeostasis. *American journal of human genetics* **78,** 179–192 (Feb. 2006).

14. Lorenz-Depiereux, B. *et al.* Hereditary hypophosphatemic rickets with hypercalciuria is caused by mutations in the sodium-phosphate cotransporter gene SLC34A3. *American journal of human genetics* **78,** 193–201 (Feb. 2006).

15. Vanakker, O. M. *et al.* Pseudoxanthoma elasticum-like phenotype with cutis laxa and multiple coagulation factor deficiency represents a separate genetic entity. *The Journal of investigative dermatology* **127,** 581–587 (Mar. 2007).

16. Ieiri, T *et al.* A 3′ splice site mutation in the thyroglobulin gene responsible for congenital goiter with hypothyroidism. *The Journal of clinical investigation* **88,** 1901–1905 (Dec. 1991).

17. Bezzina, C. R., Rook, M. B. & Wilde, A. A. Cardiac sodium channel and inherited arrhythmia syndromes. *Cardiovascular research* **49,** 257–271 (Feb. 2001).

18. Shemon, A. N. *et al.* A Thr357 to Ser polymorphism in homozygous and compound heterozygous subjects causes absent or reduced P2X7 function and impairs ATP-induced mycobacterial killing by macrophages. *The Journal of biological chemistry* **281,** 2079–2086 (Jan. 2006).

19. Akiyama, M. ABCA12 mutations and autosomal recessive congenital ichthyosis: a review of genotype/phenotype correlations and of pathogenetic concepts. *Human mutation* **31,** 1090–1096 (Oct. 2010).

20. Borg, K. *et al.* Intragenic deletion of TRIM32 in compound heterozygotes with sarcotubular myopathy/LGMD2H. *Human mutation* **30,** E831–844 (Sept. 2009).

21. Hong, J. *et al.* The novel compound heterozygous mutations, V434del and W666X, in WFS1 gene causing the Wolfram syndrome in a Chinese family. *Endocrine* **35,** 151–157 (Apr. 2009).

22. Hampson, G., Konrad, M. A. & Scoble, J. Familial hypomagnesaemia with hypercalciuria and nephrocalcinosis (FHHNC): compound heterozygous mutation in the claudin 16 (CLDN16) gene. *BMC nephrology* **9,** 12 (2008).

23. Valverde, P, Healy, E, Jackson, I, Rees, J. L. & Thody, A. J. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature genetics* **11,** 328–330 (Nov. 1995).

24. Box, N. F., Wyeth, J. R., O'Gorman, L. E., Martin, N. G. & Sturm, R. A. Characterization of melanocyte stimulating hormone receptor variant alleles in twins with red hair. *Human molecular genetics* **6,** 1891–1897 (Oct. 1997).

25. Beaumont, K. A., Shekar, S. N., Cook, A. L., Duffy, D. L. & Sturm, R. A. Red hair is the null phenotype of MC1R. *Human mutation* **29,** E88–94 (Aug. 2008).

26. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature genetics* **39,** 1443–1452 (Dec. 2007).

27. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation research* **615,** 28–56 (Feb. 2007).

28. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics* **83,** 311–321 (Sept. 2008).

29. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics* **5,** e1000384 (Feb. 2009).

30. Hofman, A, Grobbee, D. E., de Jong, P. T. & van den Ouweland, F. A. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *European journal of epidemiology* **7,** 403–422 (July 1991).

31. Hofman, A. *et al.* The Rotterdam Study: objectives and design update. *European journal of epidemiology* **22,** 819–829 (2007).

32. Hofman, A. *et al.* The Rotterdam Study: 2010 objectives and design update. *Eur J Epidemiol* **24,** 553–572 (2009).

33. Estrada, K. *et al.* A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. *Human molecular genetics* **18,** 3516–3524 (Sept. 2009).

34. Liu, F. *et al.* Digital quantification of human eye color highlights genetic association of three new loci. *PLoS genetics* **6,** e1000934 (May 2010).

35. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics (Oxford, England)* **23,** 1294–1296 (May 2007).

36. Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American journal of human genetics* **70,** 425–434 (Feb. 2002).

37. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature genetics* **37,** 413–417 (Apr. 2005).

38. 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).

39. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS biology* **8,** e1000294 (Jan. 2010).

40. Wang, K. *et al.* Interpretation of association signals and identification of causal variants from genome-wide association studies. *American journal of human genetics* **86,** 730–742 (May 2010).

41. Evans, D. M., Barrett, J. C. & Cardon, L. R. To what extent do scans of non-synonymous SNPs complement denser genome-wide association studies? *European journal of human genetics: EJHG* **16,** 718–723 (June 2008).

42. The International HapMap Project. *Nature* **426,** 789–796 (Dec. 2003).

43. Via, M., Gignoux, C. & Burchard, E. G. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome medicine* **2,** 3 (2010).

44. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5,** e1000529 (June 2009).

45. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11,** 499–511 (July 2010).

2.3 Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations

***Struchalin M*[1], *Abbas Dehghan*[1], *Jacqueline C Witteman*[1], *Cornelia M van Duijn*[1] *and Yurii S Aulchenko*[1,2]**

[1] *Department of Epidemiology, Erasmus MC, Rotterdam, 3000 CA, The Netherlands,*
[2] *Quantitative Integrative Genomics Group, Institute of Cytology and Genetics SD RAS, Novosibirsk, 630090, Russia*

Abstract

BACKGROUND: Presence of interaction between a genotype and certain factor in determination of a trait's value, it is expected that the trait's variance is increased in the group of subjects having this genotype. Thus, test of heterogeneity of variances can be used as a test to screen for potentially interacting single-nucleotide polymorphisms (SNPs). In this work, we evaluated statistical properties of variance heterogeneity analysis in respect to the detection of potentially interacting SNPs in a case when an interaction variable is unknown.
RESULTS: Through simulations, we investigated type I error for Bartlett's test, Bartlett's test with prior rank transformation of a trait to normality, and Levene's test for different genetic models. Additionally, we derived an analytical expression for power estimation. We showed that Bartlett's test has acceptable type I error in the case of trait following a normal distribution, whereas Levene's test kept nominal Type I error under all scenarios investigated. For the power of variance homogeneity test, we showed (as opposed to the power of direct test which uses information about known interacting factor) that, given the same interaction effect, the power can vary widely depending on the non-estimable direct effect of the unobserved interacting variable. Thus, for a given interaction effect, only very wide limits of power of the variance homogeneity test can be estimated. Also we applied Levene's approach to test genome-wide homogeneity of variances of the C-reactive protein in the Rotterdam Study population ($n = 5959$). In this analysis, we replicate previous results of Paré and colleagues (2010) for the SNP rs12753193 ($n = 21,799$).
CONCLUSIONS: Screening for differences in variances among genotypes of a SNP is a promising approach as a number of biologically interesting models may lead to the heterogeneity of variances. However, it should be kept in mind that the absence of variance heterogeneity for a SNP can not be interpreted as the absence of involvement of the SNP in the interaction network.

BMC Genetics,
2010 Oct 13;11:92.

Background

Genome-wide association (GWA) study has become the tool of choice for the identification of loci associated with complex traits. In GWA analysis, the association between a trait of interest and genetic variation is studied by using thousands of subjects typed for hundreds of thousands of polymorphisms. Thus several hundred loci for dozens of complex human disease and quantitative traits have been discovered utilizing this method [1].

However, it has become clear that for most complex traits, loci discovered using GWA studies currently explain a small portion of total trait's heritability and are not likely to explain all of the heritability of the trait even with additional new loci discovered using progressively larger sample sizes [2, 3]. A number of strategies that may help discovering the sources of this ”missing heritability” have been suggested [4]. In particular, it was suggested that exploring more complex genetic models, such as these accounting for gene-gene (epistatic) and gene-environment interactions is a promising approach. In the context of genetics, interactions refer to a phenomenon when the effect of an allele at a particular locus changes given the value of another (interacting) factor, which may be another allele at the same locus (e.g. dominance inter-locus interactions), or alleles at other loci (epistasis) or some other factor (end- or exogenous environment).

However detection of epistatic and gene-environment interactions is a challenging task. In GWA scans, millions of SNPs are typed and imputed [5]. Compared to standard analysis of marginal effects, a direct search for pairs of interacting loci roughly squares the number of tests to be performed making this task both computationally and methodologically difficult. A search for gene-environment interaction, unless there are *a priory* evidence that particular environmental factor is highly likely to interact with genotype, involves search of the interacting environmental factor throughout the environmental and phenomic space, again, increasing the number of tests to be performed, and leading to computational and methodological challenge.

If a method allowing detection of SNPs potentially involved in interaction networks based on the SNP and trait information (but not the information about the interacting factor(s)) existed, that would provide a substantial advancement to the field. Indeed, if such method existed, we could first screen potentially interacting SNPs using such method, and then restrict the search for the other interacting factor (genetic or environmental) to these SNPs only, dramatically decreasing the search space.

It has been suggested that analysis of equality and heterogeneity of variances of the trait between different genotypes may become such a tool [6]. If a particular genotype is interacting with some (yet unknown) factor, it could modify the marginal mean (computed from the model not including the interactor) of a trait of subjects having this genotype, and it will also increase the marginal variance of the trait: in effect the distribution of the trait in the group of subjects with

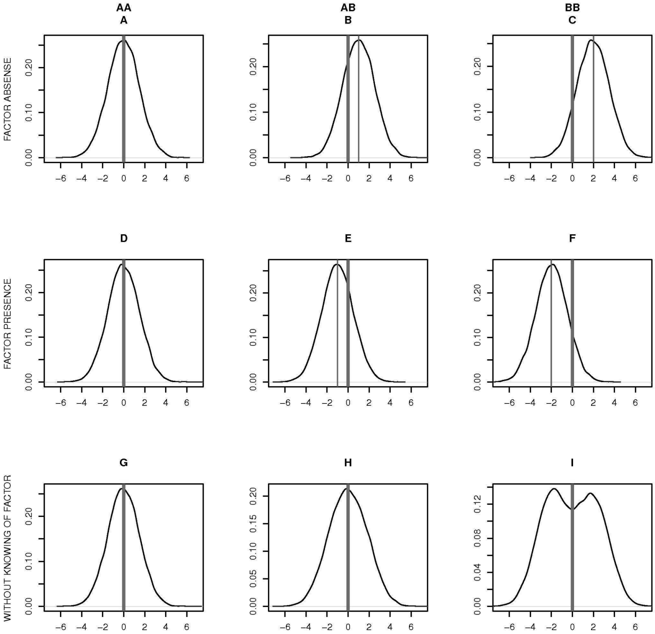


Figure 2.7: **Distribution of hypothetical trait with expectation determined by genotype and its interaction with a binary trait.** A, B, C: distribution of the trait for genotypes AA, AB, BB, correspondingly in a case when the interacting factor is present. D, E, F: distribution of the trait for genotypes AA, AB, BB, correspondingly in a case when the factor is absent. G, H, I: distribution of the trait for genotypes AA, AB, BB, correspondingly in a case when factor is unknown. In this case the distributions present mixtures of upper two ones.

interacting genotype will be described by a mixture of distributions with different means, leading to increased variance of the trait within this group.

Figure 2.7 shows the distribution of a hypothetical trait in a case of a binary factor interacting with a SNP. The upper three plots show distribution of the trait for each genotype in case of presence of the factor. Three plots in the middle show distribution of the trait for each genotype in case of absence of the factor. The lower three plots show distribution of the trait for each genotype in case when the factor is unknown and distinguishing of subjects by the factor is impossible. Theses three plots are the mixture of the distributions from upper plots for each genotypes correspondingly.

In this work, we assume an underlying model, in which the trait is generated based on knowledge of the SNP genotype and the interacting factor, and using fixed assumed model parameters. The analysis of variances of the trait is based on SNP information only, as the interacting factor is assumed to be unknown in such analysis aimed to identify potentially interacting SNPs without knowledge of an interacting variable. Using this defined framework we first evaluate type I error of different variance heterogeneity tests using simulated data. Second, assuming known interaction model involving SNP and an interacting factor, we relate the power of the variance heterogeneity test to the parameters of the underlying model.

Underlying model of the trait

We assumed the following linear model:

$$y_i \sim \mu + \beta_g g_i + \beta_F F_i + \beta_{gF} \cdot g_i F_i + \epsilon_i, \tag{2.11}$$

where y_i is a value of the trait for i^{th} individual, μ is intercept, β_g is effect of a SNP, β_F is effect of an interacting factor, β_{gF} is effect of interaction between the SNP and the factor, $g_i \sim B(n_g, P_B)$ is a SNP, which is assumed to be binomialy distributed with $n_g = 2$ (number of alleles in the genotype) and $P_B \in [0;1]$ (frequency of the interacting B allele). Below the notation AA, AB and BB is used for indicating a genotype having zero, one and two interacting alleles B correspondingly. $F_i \sim N(\mu_F, \sigma_F^2)$ is a factor, which is assumed to be normally distributed with mean μ_F and variance σ_F^2. ϵ_i is residual random error. Since many traits regularly are not normally distributed we studied seven types of distribution of ϵ_i: normal distribution, t-distribution (with $df = 2, 5, 10$) and χ^2 distributions (with $df = 1, 5, 15$). ϵ_i was standardized to have zero mean and variance of one. We assumed that the distributions of g_i, F_i, and ϵ_i are independent.

Without loss of generality we can assume that $\mu = \mu_F = 0$, and $\sigma_F^2 = 1$.

Homogeneity of variance tests

Bartlett's test is defined as:

$$T^2 = \frac{(N-k)ln(\sigma_p^2) - \sum_{j=0}^{k-1}(n_j - 1)ln(\sigma_j^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{j=0}^{k-1}\left(\frac{1}{n_j-1} - \frac{1}{N-k}\right)\right)}, \tag{2.12}$$

where k is the number of genotypes tested, n_j is the sample size of the j^{th} group (j possess the integer values from zero to $k-1$), $N = \sum_{j=0}^{k-1} n_j$ is the total sample size, $\sigma_j^2 = \frac{1}{n_j}\sum_{i=1}^{N}(y_i - \bar{y}_j)^2 I_{g_i=j}$ is variance of the j-th group, where $I_{a=b}$ is an indicator variable taking value one if $a = b$ and zero otherwise. y_i is a value of the trait for i_{th} individual, g_i is a SNP of i^{th} individual, $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{N} y_i I_{g_i=j}$ is mean value of the trait for group j, $\sigma_p^2 = \frac{1}{N-k}\sum_{j=0}^{k-1}(n_j - 1)\sigma_j^2$. Under a null hypothesis of variance homogeneity, the value of the test, T^2, is distributed as $\chi^2_{df=(k-1)}$.

Bartlett's test with prior rank-transformation to normality was done by applying Bartlett's test to a transformed trait. Rank-transformation to normality is transformation (in absence of ties) that leaves the same ranks but distribution becomes perfectly normal.

Levene's (Brown-Forsythe) test is defined as:

$$T^2 = \frac{(N-k)\sum_{j=0}^{k-1} n_j (Z_{j\cdot} - Z_{\cdot\cdot})^2}{(k-1)\sum_{i=1}^{N}\left(Z_i - Z_{g_i\cdot}\right)}, \tag{2.13}$$

where $Z_i = |y_i - \widetilde{y}_{g_i}|$, $\widetilde{y}_{g_i}$ is median value of the trait for genotype g_i, $Z_{j\cdot} = \frac{1}{n_j}\sum_{i=1}^{N} Z_i I_{g_i=j}$, $Z_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{N} Z_i$.

Under a null hypothesis of variance homogeneity, the value of the test, T^2, is distributed as F with $df_1 = (k-1)$ and $df_2 = (N-k)$ degrees of freedom. In our case, where $N = 10000$, T^2 is excellently approximated with $\chi^2_{df=(k-1)}$.

The number of genotypes is at the most three, which corresponds to genotypes AA, AB and BB. Thus, the variance homogeneity test results to a test with two degrees of freedom. We also considered three tests with one degree of freedom that test variance of a particular genotype against two others (AA vs. AB and BB, AB vs. AA and BB, and BB vs. AA and AB). For those tests we reduced trait's distribution of each genotypes to zero mean.

Simulations

To study Type I error, simulations were performed. Effects of a factor and an interaction term were set to zero ($\beta_F = \beta_{gF} = 0$). Interacting allele frequencies studied were set to 5%, 10%, 25%, and 50%. For each fixed allelic frequency, we set the effect of SNP, β_g in order to explain 0%, 1%, and 5% of the total variance of the trait. Denoting this proportion as r^2, the corresponding SNP effect was computed as

$$\beta_g = \sqrt{\frac{r^2\sigma^2_\epsilon}{(1-r^2)2P_B(1-P_B)}}, \tag{2.14}$$

where σ^2_ϵ is variance of a residual error which was assumed to be one. Thus, for each from one and two degrees of freedom tests eighty four models were studied. For each model point we simulated data for $10,000$ individuals, and simulations were repeated $10,000$ times.

Under the alternative hypothesis, assuming normally distributed residual error, we have developed an analytical expression for NCP (see subsubsection **Power** in subsection **Results**).To check correctness of our analytical solutions, we have studied several points from the model space by simulations. The parameters studied were allele frequency $P_B = \{0.05, 0.5\}$, SNP effect $\beta_g = \{0, 0.3\}$, and effect of factor $\beta_F = \{0, 1\}$.

Power of direct test for interactions

The difference in power between direct method and variance homogeneity tests were also studied. Direct test was defined as regression analysis when all variables, including the interacting factor, are known and relationships between dependent and independent variables are estimated.

Power is a function of non-centrality parameter. Analytical expression for non-centrality parameter (NCP) of test statistics to detect effect of interaction β_{gF} by direct test is

$$NCP = \beta_{gF}^2 \frac{1}{\sigma_\epsilon^2} N \left(\sigma_{gF}^2 - \frac{cov^2(F, g \cdot F)}{\sigma_F^2} \right), \tag{2.15}$$

where $\sigma_{gF}^2 = 2P_B(1 + P_B)\sigma_F^2$, $cov(F, g \cdot F) = 2P_B\sigma_F^2$, is covariance between F and $g \cdot F$.

Results

Type I error

Figure 2.8 shows type I error rate obtained in our simulation study for different variance homogeneity tests. Type I error corresponds to the threshold $\alpha = 5\%$ and interacting allele frequency 10%. Plot A shows the results for the model without SNP effect, whereas plot B represents results for the model with SNP effect explaining 5% of the total trait's variance. Each column presents one distribution of residual error, each group of columns represents one variance homogeneity test. For both figures, the interacting allele frequency $P_B = 10\%$.

From Figure 2.8, one can see that type I error of Bartlett's test grows with increase of asymmetry as well as with heavier tails of distribution.

Bartlett's test with prior rank transformation to normality has acceptable type I error 5% only in case of SNP effect absence. Only type I error of Levene's test does not show dependence on model parameters.

In case of SNP effect presence, rank transformation to normality of a trait which follows a non-normal distribution results to perfectly normally distributed trait whereas distribution of a trait for each genotype becomes distorted. Figure S4.6 shows distribution of a trait for each genotype before and after transformation in case of SNP effect presence, explaining 5% of total variance.

Results for type I error for other frequencies of interacting allele are similar to those shown in Figure 2.8. Additional file 2, Table S1, S2, and S3 present type I error in case there is SNP effect explaining correspondingly 0%, 1%, 5% of total traits's variance. Each of these tables present result for different interacting allele frequency $P_B = 5\%$, 10%, 25%, and 50%

Results for type I error for one degree of freedom tests are presented in the tables of Additional file 3, Additional file 4, and Additional file 5. The notable

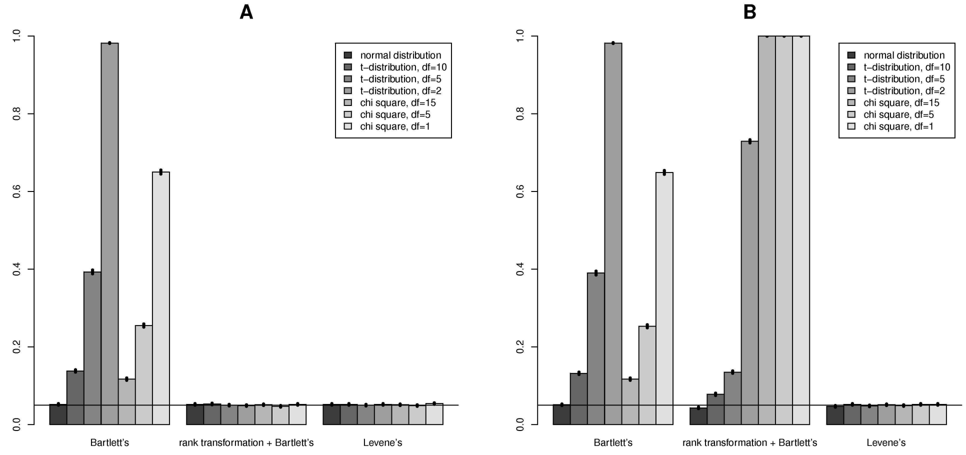


Figure 2.8: **Type I error at the threshold corresponding to $\alpha = 5\%$ for interacting allele frequency 10%. A: SNP effect is absent, B: SNP effect explains 5% of total trait's variance.**

difference from two degrees of freedom test is that even in absence of SNP effect Bartlett's test with prior rank transformation of a trait has increased type I error.

Power

We have derived an expression for dependence of trait's variances on model parameters for each genotype of a SNP.

$$\begin{aligned}
\sigma^2_{AA} &= \beta^2_F\sigma^2_F + \sigma^2_\epsilon \\
\sigma^2_{AB} &= \sigma^2_{AA} + \beta^2_{gF}\sigma^2_F + 2\beta_{gF}\beta_F\sigma^2_F \\
\sigma^2_{BB} &= \sigma^2_{AA} + 4\beta^2_{gF}\sigma^2_F + 4\beta_{gF}\beta_F\sigma^2_F,
\end{aligned}$$

where σ^2_{AA}, σ^2_{AB} and σ^2_{BB} are variances of trait's distribution in each group of subjects having corresponding genotype.

These expressions can be substituted to expression (2.12) to obtain expected NCP. These formulas were validated by simulations and results are shown in Additional file 1, Figure S2. The power to detect β_{gF} by direct test does not depend on effect of factor (F) as opposed to the homogeneity test. Figure 3 shows dependence of non-centrality parameter of variance homogeneity test on effect of factor for different frequencies of interacting allele $P_B = \{0.05, 0.4, 0.6, 0.95\}$ and different effects of interaction: the top curve on each plot shows results for interaction effect equals $\beta_{gF} = 1$, the middle curve is for $\beta_{gF} = 0.5$, and the bottom curve is for $\beta_{gF} = 0.1$.

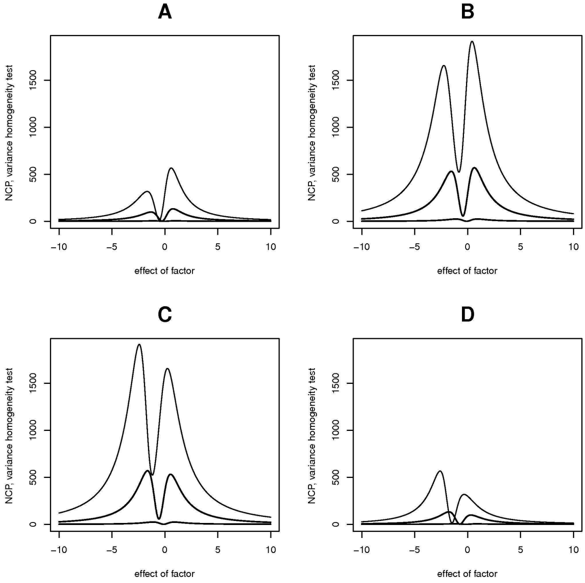


Figure 2.9: **Dependence of non-centrality parameter of variance homogeneity test on main effect of a factor.** The top curve on each plot shows results for interaction effect $\beta_{gF} = 1$, the middle curve is for $\beta_{gF} = 0.5$, and the bottom curve is for $\beta_{gF} = 0.1$. Each subplot shows different frequency of interacting allele. (A – 0.05, B – 0.4, C – 0.6, D – 0.95).

One can see that non-centrality parameter grows with increasing of interaction effect and minor allele frequency. The dependence is not monotonic and there are certain optimal effects of the factor β_F^{opt}, where the power to detect variance heterogeneity is maximum and minimum.

The plots for such dependence but for one degree of freedom tests are similar. They are shown in Additional file 1, Figures S3, S4 and S5.

It is of interest to note that NCP curves at complementary P_B (say 0.05 and 0.95) may look like mirror images at first glance: however, this symmetry is not complete. Asymmetry between plots for complementary frequencies can be explained by taking into account that heterogeneity of variances for a case $P_B^2 << 2P_B(1 - P_B)$, when genotype BB can be neglected, is determined mostly by:

$$\frac{\sigma_{AB}^2 - \sigma_{AA}^2}{\sigma_{AA}^2} = \frac{\beta_{gF}^2 + 2\beta_{gF}\beta_F}{\beta_F^2 + \frac{\sigma_\epsilon^2}{\sigma_F^2}}$$

Table 2.5: Power of variance homogeneity test under optimal effect of factor when power of direct test is 80%. Each column presents allele frequency of interacting allele, each row presents threshold α.

	5 %	40 %	60 %	95%
0.05	0.414	0.409	0.409	0.414
0.01	0.342	0.334	0.334	0.342
$5 \cdot 10^{-8}$	0.125	0.107	0.107	0.125

whereas in an opposite case, when genotype AA is neglected, heterogeneity of variances is determined by

$$\frac{\sigma_{BB}^2 - \sigma_{AB}^2}{\sigma_{AB}^2} = \frac{3\beta_{gF}^2 + 2\beta_{gF}\beta_F}{\beta_{gF}^2 + 2\beta_{gF}\beta_F + \beta_F^2 + \frac{\sigma_\epsilon^2}{\sigma_F^2}}.$$

The optimal effect of factor in the first case is given by

$$\beta_{F,AAvsAB}^{opt} = \frac{-\beta_{gF} \pm \sqrt{\beta_{gF}^2 + 4\frac{\sigma_\epsilon^2}{\sigma_F^2}}}{2}. \qquad (2.16)$$

Similarly, in second case,

$$\beta_{F,ABvsBB}^{opt} = \frac{-3\beta_{gF} \pm \sqrt{\beta_{gF}^2 + 4\frac{\sigma_\epsilon^2}{\sigma_F^2}}}{2}. \qquad (2.17)$$

Figure 2.10 shows analytical curves of dependence of power to detect interaction on effect of interaction for direct and variance homogeneity tests. Light curves present power of direct test, darker curve – upper limit of power of variance homogeneity test.

Such a dependence but for threshold corresponding to $\alpha = 5 \cdot 10^{-8}$ and $\alpha = 0.01$ is shown in Additional file 1, Figure S6.

The table 2.5 presents power of variance homogeneity test under optimal effect of factor when power of direct test is 80%. Each column presents allele frequency of interacting allele (0.05, 0.4, 0.6, 0.95), and each row presents threshold α (0.05, 0.01, $5 \cdot 10^{-8}$).

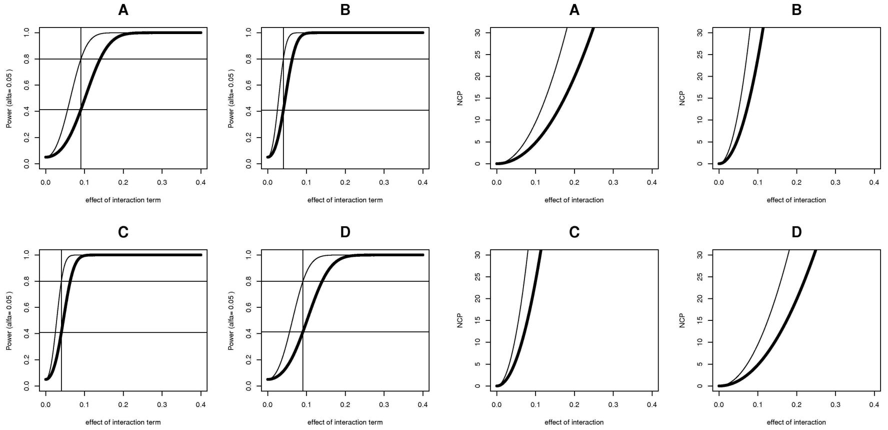


Figure 2.10: **Dependence of power to detect interaction (left plot) with threshold corresponding to $\alpha = 0.05$ and non-centrality parameter (right plot) on effect of interaction.** Thin curve on each subplot corresponds to direct test, bold curve corresponds to upper limit of variance homogeneity test. Each subplot corresponds to different frequency of interacting allele (A – 0.05, B – 0.4, C – 0.6, D – 0.95).

Performance of proposed method on real data

In order to measure the performance of the proposed method using clinical data, we applied Levene's variance homogeneity test on genome wide data for C-reactive protein (CRP), an inflammatory marker in the Rotterdam Study.

The Rotterdam Study (RS) [7] is a prospective cohort study that started in 1990 in Ommoord, a suburb of Rotterdam, and consists of 10,994 men and women aged 55 and over. The main objectives of the Rotterdam Study are to investigate prevalence, incidence and risk factors for cardiovascular, neurological, locomotor, and ophthalmologic diseases in the elderly. In the Rotterdam Study, genome-wide SNP genotyping was performed using Infinium II assay on the HumanHap550 Genotyping BeadChips (Illumina Inc., San Diego, CA, USA). In the present work, we used 5959 participants for whom genome wide and CRP data were available. Prior of applying the variance homogeneity test logarithmic transformation of CRP was performed. Genotypes from the selected SNP were tested separately. Additional file 1, Figure S7 shows genome-wide $\log(P - \text{value})$ plot and Q-Q plots respectively. Results show that no SNPs reached genome-wide significance level. The lowest $P - \text{value} = 4.77 \cdot 10^{-06}$ corresponded to SNP rs2399332 which is located on chromosome 3.

In the work of Guillaume Paré et al [6] Levene's test was applied to study CRP on a sample size of 21,799 women, and results showed a significant SNP rs12753193 located on chromosome 1 showed the lowest $P - \text{value} = 1.6 \cdot 10^{-29}$.

We tested the same SNP in Rotterdam Study and found a $P-$value of 0.011, with minor allele frequency of 0.385 for the risk-allele ″G″. The trait variances (and sample size) for genotypes AA ($n = 2098$), AG ($n = 2643$), and GG ($n = 808$) were 1.04, 1.10, and 1.18 respectively. Similarly to the work [6] genotype GG has the largest variance. From this result, we validated the genetic variant rs12753193 in the Rotterdam Study population.

Discussion

Assuming that a genotype interacts with some factor in determination of a trait's value, it is expected that the trait's variance is increased in the group of subjects having this genotype. Thus, test of heterogeneity of variances can be proposed as a test to screen for potentially interacting SNPs. In this work, we evaluated type I error and power of variance heterogeneity analysis in respect to the detection of potentially interacting SNPs under the scenario when an interaction variable is unknown.

Three different tests of variance homogeneity were chosen in order to investigate their type I error performance. They are Bartlett's, Bartlett's with prior rank-transformation to normality of a trait and Levene's (Brown-Forsythe) tests. Not surprisingly, our results were in agreement with what is known from standard statistical theory [8, 9, 10, 11]: it is known that for Bartlett's departure of the distribution of analyzed trait from normality (e.g. skewness or heavy tails) lead to increased type I error and Levene's test has better performance under these conditions. Interestingly, we have found that Bartlett's test has increased type I error even when the distribution of the trait is forced to be perfectly normal by application of rank transformation to normality in the case when the original pre-transformed distribution was non-normal, and direct effect of the SNP is present. These results, which may seem surprising at first, may be easily explained: three non-normal distributions with the same variance but different means after transformation translate to still not normal distributions with different variances. An illustrative example is provided in Additional file 1, Figure 4.6.

We showed that even if a large interaction effect is present, the power of the ″screening″ variance heterogeneity test depends strongly on the main effect of the interacting factor and may be quite limited.

This results may at first seem surprising and contra-intuitive. To help better understanding of this phenomenon, here we provide a simple example of situation when there is an interaction effect, but the variances for all genotypes are equal, thus the variance test has no power. Consider binary factor $\mathbf{F} \in \{-1, 1\}$ with effect on the trait – in accordance to our previous notation – equal to β_F, and frequency of ″1″ denoted as f (thus frequency of ″-1″ is $1 - f$). Let genotype in question to be ″dominant″ and coded as $g \in \{0, 1, 1\}$ for genotypes $\{AA, AB, BB\}$, respectively. Let mean $\mu = 0$; for simplicity, at first, let us assume that the main effect of genotype is $\beta_g = 0$. Let us denote the effect of genotype by factor interaction as β_{gF}. Let the residual variance is σ_ϵ^2. In this case, the conditional expectations of the trait for the genotype "0" are $E(y|g = 0, \mathbf{F} = -1) = -\beta_F$ (when the value of factor is -1) and $E(y|g = 0, \mathbf{F} = 1) = \beta_F$. For genotype "1", the expectations are $E(y|g = 1, \mathbf{F} = -1) = -\beta_F - \beta_{gF}$ and $E(y|g = 1, \mathbf{F} = 1) = \beta_F + \beta_{gF}$. It is easy to see that the conditional variance of the trait in genotype $g = 0$ is simply $Var(y|g = 0) = \sigma_\epsilon^2 + 4\beta_F^2 f(1 - f)$, while the variance of the trait in other genotype is $Var(y|g = 1) = \sigma_\epsilon^2 + 4(\beta_F + \beta_{gF})^2 f(1 - f)$. The conditional variances of the two genotypes are equal when either of two conditions is met: $\beta_{gF} = 0$ (absence of interaction) or $\beta_F = -\beta_{gF}/2$. Taking a simple example with $f = 1/2$ it is straightforward to see how the variance could be the same while interaction effect is present. Interestingly, if $f \neq 1/2$ and $\beta_F = -\beta_{gF}/2$, the conditional variances $Var(y|g = 0) = Var(y|g = 1)$, but conditional expectations $E(y|g = 0) \neq E(y|g = 1)$, so the interaction will translate into marginal SNP effect in the absence of the main effect (we assumed that $\beta_g = 0$). As β_F deviates from $-\beta_{gF}/2$ in any direction, the conditional variance $Var(y|g = 1)$ will increase while $Var(y|g = 0)$ will stay the same. With $|\beta_F| \to \infty$, $Var(y|g = 1) \to Var(y|g = 0)$. This explains the non-monotonic, M-shaped dependency of the non-centrality parameter of variance test on the main effect of the interaction variable demonstrated in Figure 2.

While in this work we consider a model assuming a SNP having additive effect and following Hardy-Weinberg distribution and an interaction factor following normal distribution, the same principal result – non-monotonic dependence of the power of variance test on the main effect of interacting variable – should hold for other models and other types of interacting factor (e.g. binary, as we show above, or three-level, such as other SNPs); also, a deviation from HWE will not affect our major conclusions.

Our analysis of power was performed using Bartlett's test. Bartlett's has highest power in case of normally distributed trait, but is not robust to non-normality in trait distribution. Levene's test has better performance under deviations from normality, but has lower power compared to Bartlett's test. Therefore our principal findings will not change whether Bartlett's or Levene's test is used: particular figures provided estimate maximal power, but the relation of the power to the underlying model parameters will be the same for both tests.

We considered testing for heterogeneity of variances as a screening tool for potentially interacting SNPs in the context of population-based design. It has been proposed that this testing can be more effectively done in the context of monozygotic twins or migrant studies [4]. While these designs may indeed be more powerful compared to population-based design, the same relation between power of variance heterogeneity test and the underlying model parameters is to be expected in these designs as well.

Thus, for a wide range of designs, models and test used, we can conclude that that absence of significant heterogeneity of variances can not be interpreted as absence of strong interaction because the power of the variance test depends much on the main effect of the (unobserved) interacting factor.

It is interesting to consider whether presence of significant variance heterogeneity tells us that a SNP indeed interacts with some factor. First of all, variance heterogeneity will be detected for a SNP having main effect when the distribution of the trait is heteroscedastic, i.e. the variance increases with the mean – a situation rather common in biology. This suggests that prior test for heteroscedasity should be performed before running variance heterogeneity as an ″interaction screening″ test. Another – biological – possibility is that a genotype indeed affects the variance of the trait without any specific interaction. We can speculate that there may be genotypes which affect the stability of development or homeostasis, leading to wider trait's variance.

Detection of a variance homogeneity for a given SNP does not necessary indicate that a single factor is interacting with a studied SNP. Moreover, it can suggest the presence of a complex network with many other SNPs and factors involved. The variance heterogeneity test may be especially effective to detect such SNPs – in case of multiple interacting factors it is very unlikely that the cumulative effects of the interacting factor will fall into the point at which the power of the variance test is minimal.

Further dissubsection of the SNPs demonstrating strong heterogeneity of variances may be a challenging task, requiring the search of the interactors through phenomic screening. Straightforward testing whether the identified interactor does explain heterogeneity of variances can be easily performed by using the variance homogeneity test on the residuals from the regression involving identified factor.

A number of genetic interaction models may lead to variance heterogeneity. These are straightforward interaction models as discussed above, when an environmental of other genetic factor changes the expectation of the trait value in the concert with the SNP studied. Other interesting model, leading to specific increase of the variance of the heterozygous genotype, is parent-of-origin model, when the expectation of the trait in heterozygous individuals (AB) depends on whether allele A was transmitted from father or from mother.

We showed that when one interacting factor is considered, the power of direct test, exploiting the knowledge of the interacting factor, is always greater then the power of the variance heterogeneity test. An interesting scenario in which the power of variance heterogeneity test may be greater than the power of direct test occurs when multiple interacting factors induce variance heterogeneity, in which case the power of identification any single of them (or all together) may be – due to small effects associated with particular interacting factor and with increased number of degrees of freedom – lower then the power of variance heterogeneity test.

In present GWAS, association between a SNP and a trait is studied by detecting difference between mean values of the genotypes for a given SNP. We conclude that screening for differences in variances is a promising approach as a number of biologically interesting models may lead to the heterogeneity of variances. However, it should be clearly considered that absence of variance heterogeneity for a SNP can not be interpreted as absence of involvement of the SNP into interactions network, while the presence of significant heterogeneity may be explained not only by plain interaction with some factor, but also by other biological mechanisms and statistical artifacts.

Conclusion

The method have been proposed for genome wide search of interaction between a SNP and a factor. The method is based on testing of variance homogeneity of a trait distributions in genotypes in which no knowledge of a factor is present. We have investigated type I error and power of three variance homogeneity tests (i.e. Bartlett's, Bartlett's with prior rank transformation of a trait to normality, and Levene's). Under variation of model parameters and distribution of residual errors only Levene's test kept acceptable type I error. We have obtained an analytical expression for power to detect interaction of direct test and variance homogeneity test. We also showed that the power of variance homogeneity test has lower power comparing to direct test under any model parameters when a single interacting variable is considered. As opposed to direct test, power of variance homogeneity test depends on the main effect of a factor. This dependency is non monotonic and for a given factor effect and it has its own maximums and minimums. By replicating the results of previous study [6], we demonstrate that application of the method can lead to biologically interesting, reproducible results.

BIBLIOGRAPHY

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106,** 9362–9367 (June 2009).
2. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).
3. Aulchenko, Y. S. *et al.* Predicting human height by Victorian and genomic methods. *Eur J Hum Genet* **17,** 1070–1075 (2009).
4. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
5. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10,** 387–406 (2009).
6. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet* **6,** e1000981 (2010).
7. Hofman, A. *et al.* The Rotterdam Study: 2010 objectives and design update. *Eur J Epidemiol* **24,** 553–572 (2009).
8. Levene, H. in, 278–292 (Stanford University Press, 1960).
9. Brown, M. B. & Forsythe, A. B. Robust Tests for Equality of Variances. *Journal of the American Statistical*pages (1974).
10. Bartlett, M. S. *Properties of sufficiency and statistical tests* (Proceedings of the Royal Statistical Society Series).
11. Snedecor, G. W. & Cochran, W. G. *Statistical Methods* (Iowa State University Press, 1989).

2.4 AN R PACKAGE variabel FOR GENOME-WIDE SEARCHING OF POTENTIALLY INTERACTING LOCI BY TESTING GENOTYPIC VARIANCE HETEROGENEITY

***Struchalin M*[1]*, Najaf Amin*[1]*, Paul HC Eilers*[2]*, Cornelia M van Duijn*[1] *and Yurii S Aulchenko*[1,3]**

[1] *Department of Epidemiology, Erasmus MC, Rotterdam, 3000 CA, The Netherlands*
[2] *Department of Biostatistics, Erasmus MC, Rotterdam, 3000 CA, The Netherlands,*
[3] *Recombination and Segregation laboratory, Institute of Cytology and Genetics SD RAS, Novosibirsk, 630090, Russia*

Abstract

BACKGROUND: Hundreds of new loci have been discovered by genome-wide association studies of human traits. These studies mostly focused on associations between single locus and a trait. Interactions between genes and between genes and environmental factors are of interest as they can improve our understanding of the genetic background underlying complex traits. Genome-wide testing of complex genetic models is a computationally demanding task. Moreover, testing of such models leads to multiple comparison problems that reduce the probability of new findings. Assuming that the genetic model underlying a complex trait can include hundreds of genes and environmental factors, testing of these models in genome-wide association studies represent substantial difficulties.

We and Paré with colleagues (2010) developed a method allowing to overcome such difficulties. The method is based on the fact that loci which are involved in interactions can show genotypic variance heterogeneity of a trait. Genome-wide testing of such heterogeneity can be a fast scanning approach which can point to the interacting genetic variants.

RESULTS: In this work we present a new method, SVLM, allowing for variance heterogeneity analysis of imputed genetic variation. Type I error and power of this test are investigated and contracted with these of the Levene's test. We also present an R package, `VariABEL`, implementing existing and newly developed tests.

CONCLUSIONS: Variance heterogeneity analysis is a promising method for detection of potentially interacting loci. New method and software package developed in this work will facilitate such analysis in genome-wide context.

KEYWORDS: single-nucleotide polymorphisms (SNPs), genome-wide association (GWA), gene-environment interactions (GxE), gene-gene interactions (GxG), variance heterogeneity, environmental sensitivity, `VariABEL`, the GenABEL project

BMC Genetics,
2012 Jan 24;13:4

Background

Genome-wide association studies (GWAS) have been instrumental in identifying genetic variants involved in complex diseases. In GWAS, the relation between a trait of interest and genetic variation (usually a single nuclear polymorphism — a SNP) is studied by assessing hundreds of thousands of polymorphisms in thousands of individuals. Several hundreds of loci for dozens of complex human diseases and quantitative traits have been discovered using GWAS [1].

Though GWASs were successful in finding single loci associated with a trait, complex genetic models which include many interacting loci and environmental factors are of interest as they may help finding new loci and improve our understanding of the genetics of complex traits. A search for genetic interactions by direct analysis, in which all possible genetic models are examined, meets substantial computational and methodological difficulties. When millions of SNPs are considered, which nowadays has become routine in GWAS, testing for interaction for all possible pairwise combinations of SNPs becomes cumbersome requiring parallel computations using hundreds or thousands of CPU cores. Also, a large number of models has to be tested, resulting in multiple comparison problem, which weakens the statistical power and the possibility of new findings. For instance, if a simple interaction between two SNPs is considered in analysis of one million SNPs, approximately $5 \cdot 10^{11}$ unique SNP pairs are to be tested. This amount of tests is equivalent to running a standard "direct effects only" GWAS $5 \cdot 10^5$ times.

Thus, already the simple case of interactions between only two SNPs poses serious computational challenges. However, there is no reason why biology should not be more complex, involving more than two interacting variants. In general, a trait can be determined by a complex network of multiple interacting genes and other factors, including environmental ones. Statistical modeling of such complex network of interacting factors in genome-wide context would be a big challenge both methodologically and computationally. Prior information on loci, which are likely to be involved in a trait's control (e.g. genes in pathways implicated for specific trait) can help reducing the space of models to be tested but still does not solve the problem. For example, the protein pathway involved in Alzheimer's disease incorporates hundreds of genes. Each of them may include over 25-50 SNPs.

Another approach to dissubsection of genetic interactions consist of identification of potentially interacting loci, with further search for factors which interact with these loci. For quantitative outcomes interaction of a SNP with an unknown factor can be discovered from the trait's distribution conditional on the genotype: it is expected that trait will have larger variance for an interacting genotype [2, 3]. This assumption can be tested using a variance heterogeneity test. Such testing is easily implemented and can be performed for the whole genome in a reasonable time. It also deals efficiently with the multiple comparisons problem, as the number of models to be tested in such analysis equals to the number of SNPs regardless of the complexity of the interaction model underlying the trait. In that, the variance test is similar to the regular GWAS (where the effect of a SNP on the phenotype mean is being studied). Methodologically, this approach has resemblance to the "environmental sensitivity" analysis [4, 5].

Two groups [2, 3] demonstrated that testing variance heterogeneity in GWAS is a promising approach for finding new genes involved in interactions. However the approaches proposed up until now cannot deal with imputed SNPs. Imputations are crucial for GWAS because they not only increase power in the analysis of an individual study, but also allow subsequent meta-analysis of the obtained results.

In this work we present a method extending variance heterogeneity analysis to imputed genetic data. We also develop `VariABEL` – an R package implementing variance heterogeneity tests proposed previously and developed in this work.

Implementation

Here we describe existing variance heterogeneity tests and the newly proposed test, which is suitable for the analysis of imputed genetic data (subsubsection "Variance heterogeneity tests"). Next, we describe the setup of the simulations, which were used to study statistical properties of the new test (subsubsection "Simulations") and outline the details of implementation of our software (subsection "The `VariABEL` package").

Variance heterogeneity tests

For measuring variance heterogeneity we have implemented two tests: Levene's test [6, 7] and the test where linear regression is performed on squared residual values of a trait (Squared residual Value Linear Modeling, SVLM).

Levene's (the Brown-Forsythe) test is defined as:

$$T^2 = \frac{(N-k)\sum_{j=1}^{k} n_j (Z_{j\cdot} - Z_{\cdot\cdot})^2}{(k-1)\sum_{i=1}^{N}\left(Z_i - Z_{g_i\cdot}\right)}, \tag{2.18}$$

where $Z_i = |y_i - \widetilde{y}_{g_i}|$ is the deviation of the value of the trait of i-th individual, y_i, who has genotype g_i, from the median value of the trait in individuals having that genotype, $\widetilde{y}_{g_i}$; N is the total sample size, n_j is the number of individuals with genotype j, k is the number of possible genotypes, $Z_{j\cdot} = \frac{1}{n_j}\sum_{i=1}^{N} Z_i I_{g_i=j}$ is mean deviation from the median for individuals having genotype j ($I_{g_i=j}$ is an indicator variable which takes value of one if g_i is equal to j and zero otherwise), and $Z_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ is the mean deviation from the median across all individuals.

Under the null hypothesis of variance homogeneity, the value of the test statistic, T^2, has an F distribution with $df_1 = (k-1)$ and $df_2 = (N-k)$ degrees of freedom. In a case of large N, T^2 is approximated well by the $\chi^2_{df=(k-1)}$ distribution. With three possible genotypes, $k-1=2$.

Genetic imputations routinely used in GWAS nowadays increase the power in the analysis of individual studies and also allow meta-analysis of the studies using different SNP arrays. In case of imputations the posterior probability of a genotype is estimated for each subject for a given SNP. Because standard variance heterogeneity tests assume that an observation should be known to belong to a certain group (i.e. an individual is known to have specific genotype with full confidence), they can not be directly applied to the imputed data.

To allow for variance heterogeneity test for imputed SNPs we propose a simple procedure (SVLM) described below. It is known from elementary statistics that by definition the variance is:

$$\text{Var}(Y) = E[(Y - E[Y])^2] = E[Y^2] - E[Y]^2 \tag{2.19}$$

where Y is a random variable, $\text{Var}(Y)$ is the variance of Y, $E[Y]$ and $E[Y^2]$ are expected values of the variable Y and Y^2 correspondingly. In our case Y is a trait. The variance of Y conditional on the genotype g is $V(Y|g) = E[(Y - E[Y|g])^2|g]$. This means that for each genotype the variance is equal to the mean of the squared residual of the trait conditional on the genotype.

To explain this idea we provide Figure 2.11. Panel 2.11A shows the relation between the trait value and the number of B alleles in the genotype. It is assumed that allele B is interacting with some quantitative factor, hence the variance of the trait is increasing as the number of B alleles, present in an individual's genotype, increases. Figure 2.11B shows the same data, but the points correspond to the squared residuals after subtracting genotypic mean from the trait's value. The means of these squared residuals in each genotypic group shown in panel B is equal to the variance within genotypic groups shown in panel A.

Thus, taking squared residuals conditional on the genotype changes the task of estimation of the conditional variances into the task of estimation of the conditional means, which can be approached with using conventional methods such as regression analysis. Important covariates having large effects on means, can be easily accommodated in the model if necessary by modifying the expression used to compute the conditional mean.

Technically, the SVLM method consists of two steps. First, a regression analysis is applied where the trait is adjusted for a possible SNP effect and other covariates. Second, a regression analysis is applied to the squared values of residuals obtained from the first stage, using the SNP as the predictor.

Simulations

To study Type I error and power of the SVLM test, we performed a simulations study. Similar to our previous work [3], we simulated the trait under following linear model

$$y_i = \mu + \beta_g g_i + \beta_F F_i + \beta_{gF} \cdot g_i F_i + \epsilon_i, \tag{2.20}$$

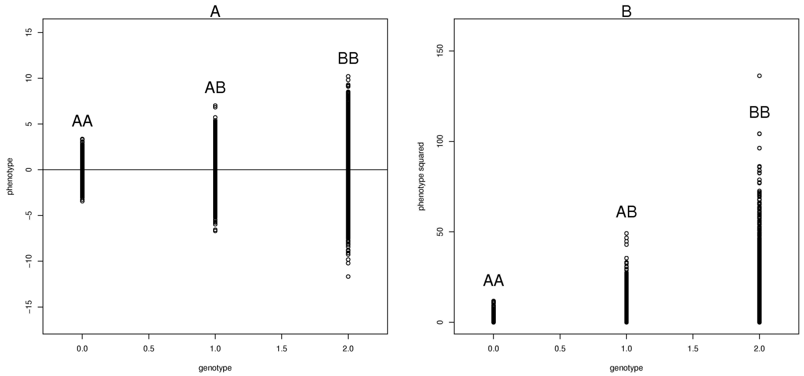


Figure 2.11: **The figure shows additive dependence of a trait (subplot A) and squared values of this trait (subplot B) on the number of alleles** ***B*** **which is in interaction with an unknown factor.** *AA*, *AB*, and *BB* stand for genotypes of biallelic SNPs. Squaring of the trait results into genotypic mean difference in case of interaction presence.

where y_i is the value of the trait for i^{th} individual, μ is the intercept, β_g is the direct effect of the SNP, β_F is the direct effect of the interacting factor, β_{gF} is the effect of interaction between the SNP and the factor, $g_i \sim B(n_g, P_B)$ is a SNP, which is assumed to be binomially distributed with $n_g = 2$ (number of alleles in the genotype) and $P_B \in [0; 1]$ (frequency of the interacting B allele). $F_i \sim N(\mu_F, \sigma_F^2)$ is a factor, which is assumed to be normally distributed with mean μ_F and variance σ_F^2. ϵ_i is the random error, which follows a normal distribution with a zero mean and a variance of one. We assumed that the distributions of g_i, F_i, and ϵ_i are independent. For our simulations, without loss of generality we can assume that $\mu = \mu_F = 0$, and $\sigma_F^2 = 1$.

Without loss of generality for both type I error and power the SNP effect was set to zero $\beta_g = 0$. We studied four different frequencies of the interacting allele: 5%, 40%, 60% and 95%. Results for 40% are presented in the text below. Results for other frequencies are shown in Supplementary Figure 4.13, Supplementary Figure 4.14 and Supplementary Figure 4.14. From the GWAS it is known that the regression analysis may lead to spurious results when the frequency of the minor allele is very low. Therefore additionally we have studied type I error of the SVLM test for allele frequencies 0.0005, 0.00075, 0.001, 0.002, 0.003, 0.004 and 0.005. As SVLM requires squaring of the trait's residuals, the presence of extreme values can affect the type I error and power of this test. To check this, we studied three types of distribution of the residual error term ϵ_i: one normal distribution and two types of χ^2 distributions with degrees of freedom $df = 1$ and $df = 5$ respectively. We simulated data for 10000 individuals.

For studying Type I error effect of the factor and the effect of interaction term were both set to zero ($\beta_F = \beta_{gF} = 0$). Twenty thousands simulations were performed.

For studying dependence between power and the interaction effect, 1000 simulations were done under each simulation scenario. As we demonstrated previously [3], the magnitude of the genotypic variance difference and hence the power of the variance heterogeneity test depends not only on the effect of interaction between the genotype and the environmental factor, but also on the magnitude of the main effect of the interacting factor F. This dependence is not monotonic and, given other parameters are fixed, there is a certain optimal main effect of the factor under which the magnitude of variance difference and, therefore, the power to detect interaction is maximal. The value of the optimal effect depends on the interaction effect, variance of the factor and the variance of error term. As in our study variance of the factor and the variance of error term is fixed to one, the optimal effect of the factor depends on the interaction effect only. For simplicity, power was studied using the optimal main effect of the interacting factor. The range of the optimal effects of the factor used in this study can be found in the Supplementary Figure 4.16.

In both type I error and power estimation, the null hypothesis was rejected when threshold p-value ≤ 0.05 was reached.

The `VariABEL` package

The `VariABEL` software implementing the SVLM is designed as an R package written in C++ and R languages. For regression analysis used by the SVLM method, the LAPACK functions ″dgeqrf″ and ″ch2inv″, which are part of the R distribution, were used. The package was compiled with gcc version 4.1.2 under Linux with version 2.6.18-274.7.1.el5 (Red Hat 4.1.2-51) and tested in R of version 2.13.1. The package is distributed under the GNU GPL license (v. 2.0 or later).

Stable version of the `VariABEL` package can be downloaded from the Comprehensive R Archive Network, CRAN [8] (`http://www.r-project.org/`). Installation is possible from R directly by running the command
″`install.packages("VariABEL")`″. Documentation is available as a part of the distribution and also on-line at the GenABEL project web-site (`http://www.genabel.org`). Developmental version of the package is available from the GenABEL project development pages (`http://genabel.r-forge.r-project.org`) located at R-forge [9].

The first stage of SVLM analysis consists of standard regression analysis which is used to access association between mean values of the trait and SNPs. `VariABEL` output contains results from both stages of analysis (modeling of means and variances). Thus, the `VariABEL` can be used for regular GWAS as well.

Results and discussion

Type I error and power

As it was mentioned above, we studied three different distributions of ϵ_i. The SVLM test had acceptable type I error for all of them: $\alpha_{\text{normal}} = 0.0471 \pm 0.0015$, $\alpha_{\chi^2_{df=5}} = 0.0488 \pm 0.00152$, and $\alpha_{\chi^2_{df=1}} = 0.04955 \pm 0.00153$ under fixed threshold $p \leq 0.05$. We did not see any significant deviation from nominal type I error rate of 5% for allele frequencies 5%, 40%, 60% and 95%. To understand the minimum sample size in a genotypic group under which the type I error of SVLM test still stays at a nominal level we measured type I error for allele frequencies 0.0005, 0.00075, 0.001, 0.002, 0.003, 0.004 and 0.005. These allele frequencies correspond to number of heterozygotes in a sample 10, 15, 20, 40, 60, 80 and 100. For those frequencies the type I errors (with its standard errors) were 0.028 ± 0.001, 0.032 ± 0.001, 0.037 ± 0.001, 0.042 ± 0.001, 0.044 ± 0.001, 0.049 ± 0.002 and 0.045 ± 0.002 correspondingly. This suggests that SVLM test has correct type I error rate when sample size in one of a genotypic group is not less then 80, while for smaller values the SVLM test starts being conservative. Levene's test did not show significant deviation from nominal value of 5% under these extremely low sample sizes.

Figure 2.12 shows the dependence of the power of the SVLM (triangles) and the Levene's (circles) tests on the effect of interaction for differently distributed error term (ϵ_i). Figure 2.12 shows that the power of the SVLM test depends on the skewness of the error term distribution stronger than the power of the Levene's test. When error term follows Normal distribution, the power of SVLM test is greater than the power of Levene's test (Figure 2.12, panel A). When the error term follows χ^2 distribution with $df = 5$, the power of the SVLM test and Levene's test are similar (Figure 2.12, panel B). In case of higher skewness the SVLM test has lower power than the Levene's test (Figure 2.12, panel C). This can be explained by the fact that Levene's test is known to be robust to the deviations from normality, while the SVLM test is in fact a regression analysis for which the outcome is supposed to follow a normal distribution.

Supplementary Figure 4.13, Supplementary Figure 4.14 and Supplementary Figure 4.15 show the dependence of the power of the SVLM test and the Levene's test on the effect of interaction for different frequencies of the interacting allele.

As it is expected the power of both tests decreases when allele frequency decreases. The observation that SVLM's test power is affected by skewness more than the power of Levene's test stays true for all studied allele frequencies.

Performance

The analysis by the SVLM test of 2543887 SNPs of 2715 subjects takes 46 minutes on one core of a Sun Fire X4540 Server with Quad-Core AMD Opteron Processor 2356.

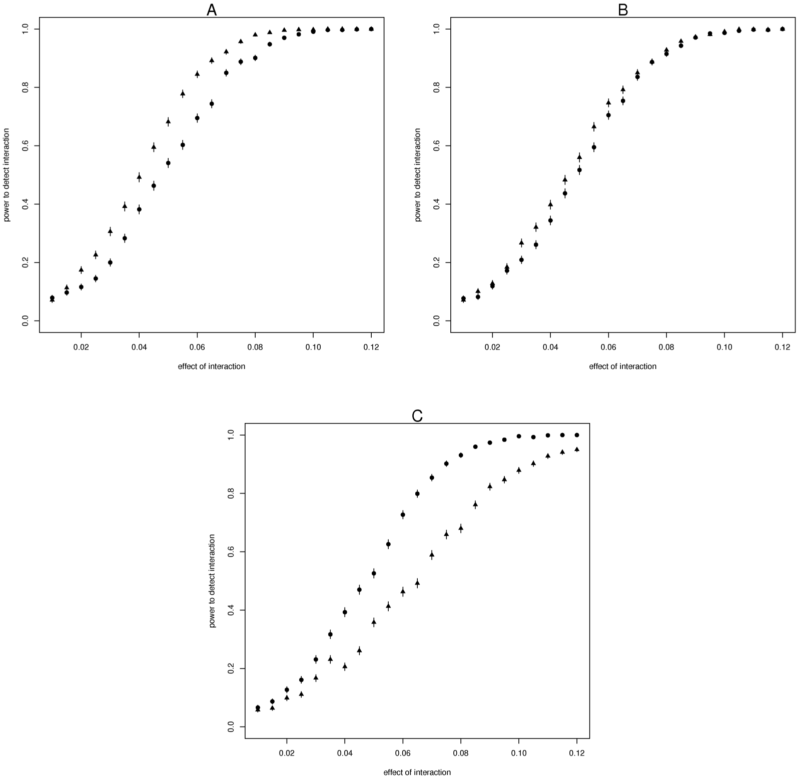


Figure 2.12: **Dependence of power to detect interaction on effect of interaction for Levene's (circles) and SVLM (triangles) tests for three different types of distribution of the error term ϵ_i: normal distribution (panel A), $\chi^2_{df=1}$ (panel B), and $\chi^2_{df=5}$ (panel C).**

Discussion

Genome-wide association analysis is currently a primary tool for identification of loci associated with complex human traits. Testing for association under complex genetic models involving multiple interactions represents methodologically and computationally challenging task.

We and others have developed a method allowing testing of SNPs genome-widely for possible involvement into interaction [3, 2] via testing of the heterogeneity of variance of the trait conditional on the genotype. Here we extend this method to imputed SNPs. The method we suggest, SVLM, is based on linear regression, and therefore results obtained in individual studies can be easily meta-analyzed using conventional methods and software tools.

Analysis of genotypic variances can be of interest to medical research. Assuming that there is a certain genotype associated with high variance of, for instance, blood pressure, the subjects having this genotype can be at risk of having extremely low or extremely high blood pressure.

In developing our method for analysis of variances using imputed data we have utilized the fact that the variance is, by definition, the expectation of squared values of the variable in case of zero mathematical expectation of this variable. This allowed us re-formulate the task of estimation and analysis of variances of the trait as a task of regression analysis of transformed trait. In this setting, methodological and computational tools developed for GWAS are applicable for the variance analysis.

The most important advantage of the proposed method is the possibility to detect SNPs belonging to a complex genetic network with many interacting factors that is impossible to study with standard tools. These SNPs will show variance heterogeneity and using our method these SNPs can be detected without knowing all the factors involved into this network. To find the factors, which interact with the identified SNP, a follow-up analysis can be applied where interaction between the SNPs found in variance analysis and all other measured SNPs or environmental factors are tested. In a case of interaction with an unknown factor, the SNPs showing significant variance differences still can be used to improve the variance explained as shown in the example below.

Consider a scenario in which SNPs, associated with a trait found in regular GWAS's, together explain a certain proportion of total trait's variance:

$$R^2_{total} = \frac{\sigma^2_{GWAS}}{\sigma^2_{total}}, \tag{2.21}$$

where R^2_{total} is the proportion of total explained variance, σ^2_{GWAS} is the variance explained by GWAS SNPs, and σ^2_{total} is the trait's variance. In addition a SNP has been found by variance analysis, showing different genotypic variances in a way where presence of interacting allele B increases trait's variance: $\sigma^2_{AA} < \sigma^2_{AB} < \sigma^2_{BB}$, where σ^2_{AA}, σ^2_{AB}, and σ^2_{BB} are variances for the respective genotypes group AA, AB, and BB. Assume that allele frequencies and the effects of the SNPs found in GWAS and which contributed into σ^2_{GWAS} are the same in each genotypic group AA, AB, and BB of the interacting SNP. Then the proportions of explained variance for different genotypic groups are:

$$\begin{aligned} R^2_{AA} &= \tfrac{\sigma^2_{GWAS}}{\sigma^2_{AA}} \\ R^2_{AB} &= \tfrac{\sigma^2_{GWAS}}{\sigma^2_{AB}} \\ R^2_{BB} &= \tfrac{\sigma^2_{GWAS}}{\sigma^2_{BB}} \end{aligned}$$

where R^2_{AA}, R^2_{AB}, R^2_{BB} are proportions of variances explained by GWAS SNPs in individuals with genotypes AA, AB, and BB, respectively, at the SNP identified by the variance analysis. Taking into account that $\sigma^2_{AA} < \sigma^2_{AB} < \sigma^2_{BB}$ it follows that the proportions explained variance by the GWAS SNPs is higher in genotypic group AA compared to AB and BB, and higher in genotypic group AB compared to BB: $R^2_{AA} > R^2_{AB} > R^2_{BB}$. The value of the proportion of total explained variance (R^2_{total}) is between R^2_{AA} and R^2_{BB} and this value depends on interacting allele frequency, effect of interaction, variance and effect of interacting factor. Thus, in such a scenario there is at least one genotypic group (AA) for which SNPs found in GWAS's explain more of the trait's variance σ^2_{AA} compared to the total trait variance σ^2_{total}.

To perform genotypic variance analysis for pedigree-based studies we propose to use GRAMMAR [10] implemented into GenABEL software [11]. In GRAMMAR the mixed model is applied where the trait is adjusted on random additive polygenic effect. Residuals from this model are free from polygenic familiar correlations and can be used for variance analysis.

To increase power of variance analysis by including the data from other studies the same approach as for regular GWAS can be used where the analysis is done for each cohort separately, followed by meta-analysis.

The SVLM method can be used for discovering interacting SNPs following any of additive, dominant, recessive, over-dominant (where trait's variance among heterozygotes is increased), or genotypic models. In case of testing the additive variance model only, the SVLM test has maximal power in the case when the SNP follows true additive model and less power in case of dominant, recessive and over-dominant models. It is of interest to note that in case of over dominant model the power to detect interaction by the SVLM test is zero if the minor allele frequency (MAF) is 0.5 and increases with decreasing MAF. In a case when MAF is close to 0.5 Levene's test has higher performance.

Conclusion

In this work we present further development of the method for detection of potentially interacting SNPs, extending it to the case of analysis of imputed SNPs. The method is based on testing of heterogeneity of trait's variance conditional on the genotype of locus being tested. We also present an R package, `VariABEL`, to facilitate for such analysis in genome-wide context. The package implements already existing variance heterogeneity tests, and the SVLM test developed in this work.

Availability and requirements

PROJECT NAME: `VariABEL` package

PROJECT HOME PAGE: `http://www.genabel.org/packages/VariABEL`

OPERATING SYSTEMS: Linux, Mac OS X, Windows programming language: R, C++

other requirements: R ($\geq$ 2.13.0)

license: GNU GPL ($\geq$ 2)

any restrictions to use by non-academics: none except these posed by the license

List of abbreviations

GWAS: Genome-Wide Association Study; MAF: Minor Allele Frequency; SNP: Single Nuclear Polymorphism, SVLM: Squared residual Value Linear Modeling.

Competing interests

None declared

Authors' contributions

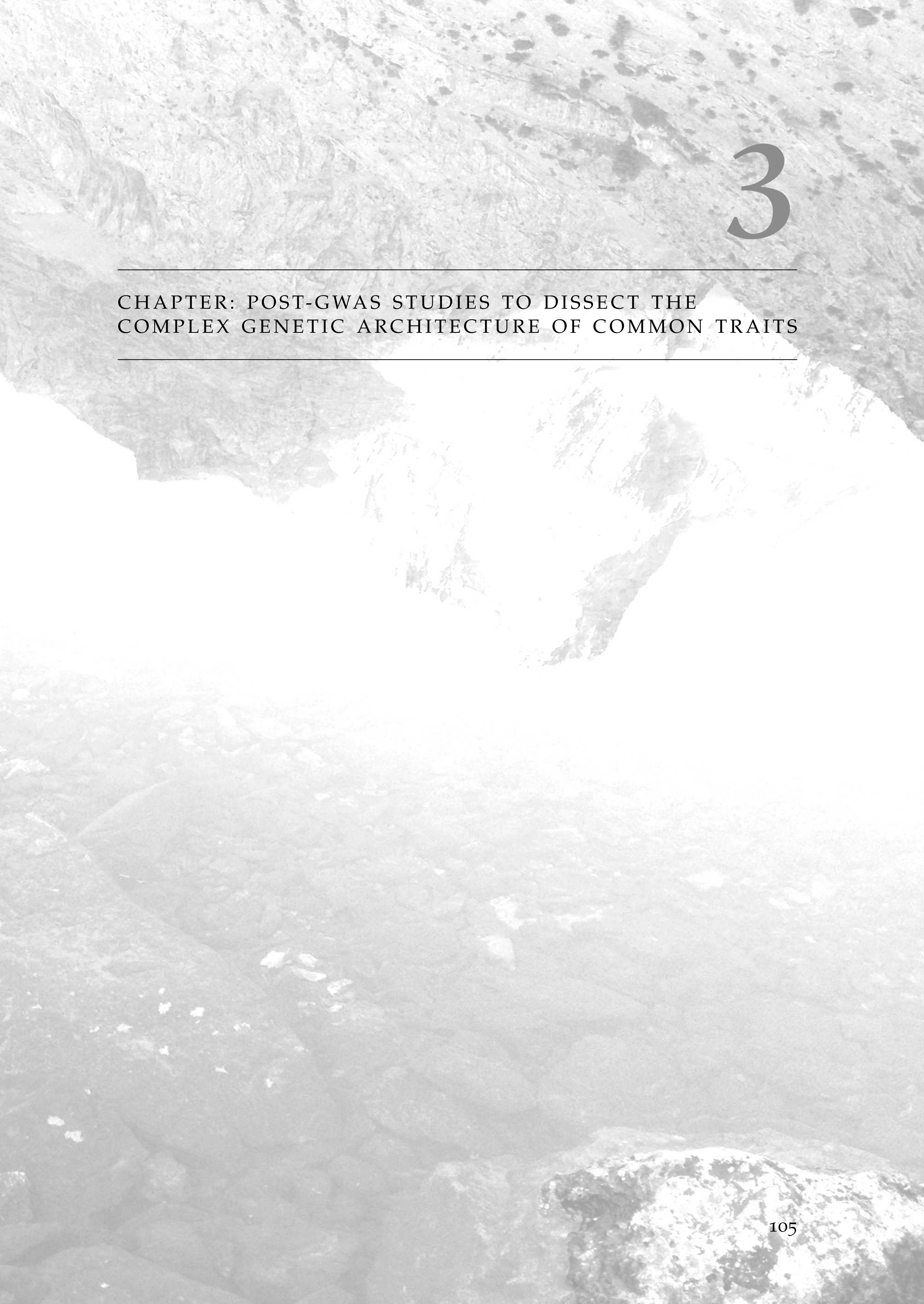
MS wrote the software, planned and carried out the simulation study and wrote the manuscript. NA, PE, CvD and YA planned the simulation study and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank Ayse Demirkan, Lennart Karssen, Xia Shen, and Anastasija Zvirbule for their help in conducting this work. The work of YA was supported by grants from RFBR and RFBR-Helmholtz JRG.

BIBLIOGRAPHY

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106,** 9362–9367 (June 2009).
2. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet* **6,** e1000981 (2010).
3. Struchalin, M. V., Dehghan, A., van Duijn, J. C. W. C. & Aulchenko, Y. S. Variance heterogeneity analysis for detection of potentially interactinggenetic loci: method and its limitations. *BMC Genet* **11,** 92 (2010).
4. Falconer, D. Selection in different environments: effects on environmental sensitivity (reaction norm) and on mean performance. *Genet Res* **56,** 57–70 (1990).
5. Visscher, P. M. & Posthuma, D. Statistical power to detect genetic Loci affecting environmental sensitivity. *Behav Genet* **40,** 728–733 (2010).
6. Levene, H. in, 278–292 (Stanford University Press, 1960).
7. *The car R project* <http://cran.r-project.org/web/packages/car/>.
8. *CRAN* <http://cran.r-project.org/web/packages/>.
9. *R-forge* <https://r-forge.r-project.org/R/?group_id=505>.
10. Aulchenko, Y. S., de Koning, D.-J. & Haley, C. Genomewide rapid association using mixed model and regression: afast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177,** 577–585 (2007).
11. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).



3

Chapter: Post-GWAS studies to dissect the complex genetic architecture of common traits

3.1 PREDICTING HUMAN HEIGHT BY VICTORIAN AND GENOMIC METHODS

***Yurii S Aulchenko*[1,2,7]*, Struchalin M*[1,3,7]*, Nadezhda M Belonogova*[2,4]*, Tatiana I Axenovich*[2]*, Michael N Weedon*[5]*, Albert Hofman*[1]*, Andre G Uitterlinden*[6]*, Manfred Kayser*[3]*, Ben A Oostra*[1]*, Cornelia M van Duijn*[1]*, A Cecile JW Janssens*[1] *and Pavel M Borodin*[2,4]**

[1] *Department of Epidemiology and Biostatistics and Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands;*
[2] *Laboratory of Recombination and Segregation Analysis, Institute of Cytology and Genetics SD RAS, Novosibirsk,Russia;*
[3] *Department of Forensic Molecular Biology, Erasmus MC, Rotterdam, The Netherlands;*
[4] *Department of Cytology and Genetics, Novosibirsk State University, Novosibirsk, Russia;*
[5] *Department of Genetics of Complex Traits and Diabetes Genetics, Peninsula College of Medicine and Dentistry, Exeter, UK;*
[6] *Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands;*
[7] *These authors contributed equally to the work.*

Abstract

In the Victorian era, Sir Francis Galton demonstrated that "when dealing with the transmission of stature from parents to children, the average height of the two parents, ... is all we need care to know about them" (1886). One hundred twenty two years after Galton's work was published, 54 loci showing strong statistical evidence for association to human height were described, providing us with potential genomic means of human height prediction. In a population-based study of 5748 people we find that 54-loci genomic profile explained 4-6% of the sex- and age-adjusted height variance, and had limited ability to discriminate tall/short people, as characterized by the Area Under the receiver-operating characteristic Curve (AUC). In a family based study of 550 people with both parents having height measurements we find that the Galtonian mid-parental prediction method explained 40% of the sex- and age-adjusted height variance and demonstrated high discriminative accuracy. We have also explored how much variance a genomic profile should explain to reach certain AUC values. For highly heritable traits such as height, we conclude that in applications where parental phenotypic information is available (e.g. medicine), the Victorian Galton's method will long stay unsurpassed, both in terms of discriminative accuracy and costs. For less heritable traits and in situations when parental information is not available (e.g. forensics), genomic methods may provide an alternative, given the variants determining an essential proportion of trait.s variation could be identified.

Eur J Hum Genet.,
2009 Aug;17(8):1070-5

Height is a classical example of an inherited human trait. More then 100 years ago, Francis Galton used height data to study resemblance between parents and offspring concluding that "when dealing with the transmission of stature from parents to children, the average height of the two parents, .. is all we need care to know about them [1] (Figure 3.1). Later on height was among the first phenotypes studied using the polygenic model of inheritance [2] which bridged the gap between the Galtonian and Mendelian genetics. Numerous studies following the pioneering work of Galton demonstrated that height is one of the most heritable human phenotypes. Typically, the proportion of the sex- and age-adjusted variance of height attributable to familial factors (heritability) is estimated as 80%. Most of this heritability may be due to genetic factors because for height the non-genetic causes of sib resemblance are usually negligibly small [3]. Up until recently, however, little was known about the genes involved in normal variation of height in human populations.

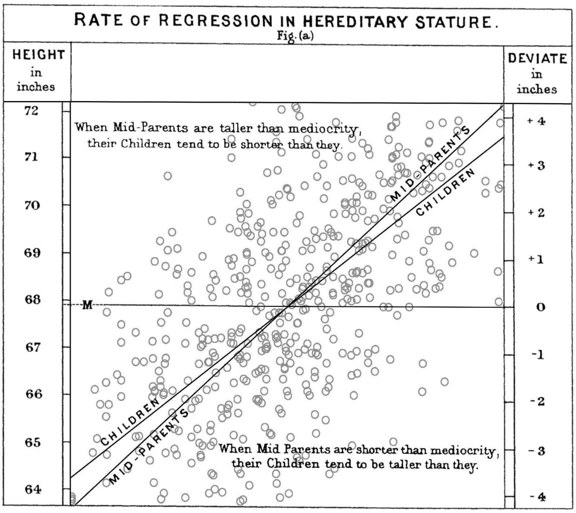


Figure 3.1: **Rate of regression in hereditary stature (Plate IX, figure a from Galton [1] with superimposed data from the ERF study).**

One hundred and twenty two years after Galton's paper, and seven years after the initial sequencing of the human genome [4], three manuscripts described 54 loci showing strong statistical evidence for association with height [5, 6, 7], potentially providing us with genomic means of human height prediction. Here, we investigate the potential of the state-of-the-art genomic approach to predict human height and compare it to the potential of 122-year-old Victorian method of Galton.

Material and methods

Study populations

The Rotterdam Study [8] is a prospective cohort study that started in 1990 in Ommoord, a suburb of Rotterdam, among 10, 994 men and women aged 55 and over. The main objective of the Rotterdam Study is to investigate the prevalence and incidence and risk factors for cardiovascular, neurological, locomotor and ophthalmologic diseases in the elderly. Baseline measurements were obtained between 1990 and 1993. All participants were subsequently examined in follow-up examination rounds every 2-3 years. Height measurements were performed at baseline. The Rotterdam Study has been approved by the institutional review board (Medical Ethics Committee) of the Erasmus Medical Center and by the review board of the Netherlands Ministry of Health, Welfare and Sports. For this study, we used data on 5,748 participants for whom GWA and height data were available. Erasmus Rucphen Family (ERF) study [9] is a family based study of a young genetically isolated population studied within Genetic Research in Isolated Populations programme [10]. ERF study includes over 3000 participants descending from 22 couples living in the Rucphen region in the 19th century. All descendants were invited to visit the clinical research center in the region where they were examined in person, including height measurements. ERF study has been approved by the Medical Ethics Committee of the Erasmus MC. In this study we included 550 participants together with both parents for whom height measurements were complete.

Genotyping and imputations in the Rotterdam Study

In Rotterdam Study, genome-wide SNP genotyping was performed using Infinium II assay on the HumanHap550 Genotyping BeadChips (Illumina Inc, San Diego, USA). Approximately 2.5 million SNPs were imputed using HapMap CEU population (release 22) as reference. The imputations were performed using MACH software [11].

The quality of imputations were checked by contracting imputed and actual genotypes at 78, 844 SNPs not present on Illumina 550K for 437 individuals for whom these SNPs were directly typed using Affymetrix 500K. Using the "best guess" genotype for imputed SNPs the concordance rate was 99% for SNPs with R^2 (ratio of the variance of imputed genotypes to the binomial variance) quality measure greater than 0.9; concordance was still high (94%) when R^2 was between 0.5 and 0.9. Out of 54 SNPs used in this study, 31 were directly typed and the rest were imputed. The median R^2 was 0.999 and only two SNPs had $0.87 < R^2 < 0.9$ (Supplementary Table 4.3). In ERF study, genome-wide SNP genotyping was performed using Illumina HumanHap300 (1, 200 individuals), HumanHap370 (100 individuals) and Affymetrix 250K Nsp array ($\sim$ 200 individuals). The imputations followed the Rotterdam study protocol closely.

Selection of 54 SNPs used in the study

Of 54 loci influencing human height shown in Supplementary Table 4.3, 16 are published by Weedon et al. [5], 11 by Lettre et al. [6] and 27 by Gudbjartsson et al. [7]. Of 59 markers reported to be strongly associated with height in these three studies, five were mapped within the same chromosome region. For these loci we picked up markers with lowest p-value.

Table 3.1: **Proportion of human height variance explained and discriminative accuracy of different predictive profiles**

				AUC discriminating			
Profile	Pop	N	σ^2,%	Top 50%	Top 5%	Top 1%	$\Delta_{5,95}$, cm
54-loci genomic	RS	5748	3.8	58.1	64.8	63.4	4.95
Hypothetical	RS	5748	80.0	93.3 ± 0.2	97.4 ± 0.3	98.8 ± 0.2	23.4 ± 0.01
Galtonian mid-parental	ERF	550	40.1	77.3	83.6	97.4	17.68
Galtonian mid-parental[a]	ERF	257	44.9	78.7	88.3	99.9	21.18
Galtonian+54-loci[a]	ERF	257	46.2	80.0	88.9	98.2	21.28

σ^2 - variance explained;
AUC - area under the receiver-operating characteristic curve;
$\Delta_{5,95}$ - difference between mean height of people coming from the top 5% and bottom 5% of the profile distribution;
RS - population-based Rotterdam Study;
ERF - Erasmus Rucphen Family study.
a - ERF participants with parental and genotypic information ($n = 257$).

Testing within- and between-loci additivity

All analyses were performed using R v 2.7.0 (`http://www.r-project.org`). To test deviation from within-locus additive model, we used linear model $height \sim \beta_s sex + \beta_a age + \beta_{AB} P_{AB} + \beta_{BB} P_{BB}$, where P_{AB} and P_{BB} are the estimated probabilities of the AB and BB genotypes, respectively. This model was contrasted to the model under additive restriction $\beta_{BB} = 2\beta_{AB}$ using Likelihood Ratio Test, LRT (twice the difference between maximum log-likelihood of these models is asymptotically distributed as χ^2_1). Multiple testing was accounted for using Bonferroni correction. Similarly, we have tested the deviation from between-loci additivity using the model $height \sim \beta_s sex + \beta_a age + \beta_{S1} D_{S1} + \beta_{S2} D_{S2} + \beta_I D_{S1} D_{S2}$, where D_{S1} and D_{S2} were the estimated allele doses at two loci ($D = P_{AB} + 2P_{BB}$) and β_I is the interaction term. This model was contrasted to the no interaction model ($height \sim \beta_s sex + \beta_a age + \beta_{S1} D_{S1} + \beta_{S2} D_{S2}$); again, LRT on one degree of freedom was performed for comparison.

Construction of predictive profiles

The non-weighted allelic profile was computed as the sum of the estimated doses of the height-increasing allele in the genotype of a person. The weighted allelic profile was constructed as a weighted allelic sum with weight proportional to the allelic effect estimated using our data in a multivariable model including all 54 SNPs. To construct the Galtonian mid-parental profile, we first estimated height residuals from the model $height \sim sex + age$. For every person for whom both paternal and maternal height was available, we then constructed the "predictive profile" which was defined as the average of the parental height residuals. This methods resembles very closely the method of Galton [1], with the exception that he did not adjust for age. The hypothetical predictor explaining a certain proportion (V_e) of sex- and age-adjusted height variance was constructed as a sum of person's height plus a Normally distributed random number, with mean zero and variance equal to $(V_h \cdot (1 - V_e)/V_e)$, where V_h is the variance of height. In any analyses involving simulated profile, we used at least 100 simulations per point of interest.

Estimating proportion of variance explained by a profile and discriminative accuracy (AUC)

The proportion of the variance of sex- and age-adjusted height explained by a profile was estimated using linear regression model as $(1 - V_i/V_e)$, where V_i is the trait's variance in the model including, and V_e is the variance in the model excluding the profile as a predictor. The ROC curve presents the combinations of sensitivity and specificity for each possible cut-off value of the continuous test result that can be considered to define positive and negative test outcomes. The area under the ROC curve (AUC) indicates the discriminative accuracy of a continuous test [12]. The AUC ranges from 0.5 (total lack of discrimination) to 1.0 (perfect discrimination) and is independent of the prevalence of the condition of interest [13]. The AUC basically can be considered as the probability that the test correctly identifies the subject possessing the characteristics of interest (e.g. "very tall") from a pair of whom one has and one has not this characteristic. An AUC of 0.95 means that 95% of the pairs is correctly classified, whereas a test with an AUC is 0.50 is non-discriminative - as accurate as tossing a fair coin. AUC was computed as the area under the function relating sensitivity to (1-specificity) (ROC curve). To derive the ROC curve, we varied the threshold determining "positive test result" from minimal to maximal possible test (profile) value. At given threshold, sensitivity was computed as the proportion of people who test positive among these who do indeed possess the characteristic of interest; the specificity was computed as the proportion of these who test negative among these who do non possess the characteristic of interest.

In all analyses, we used sex- and age-adjusted height as an outcome. We have compared the predictive potential of different methods by contrasting the proportion of height variance explained and the Area Under the receiver-operating characteristic Curve (AUC). The latter measures the accuracy of the model to discriminate between alternative outcomes (in height context, e.g. "very tall" or not).

The data from population-based Rotterdam Study [8] (5748 individuals with complete height, sex, age and genomic data) were used to estimate predictive potential of the genomic method. In the Rotterdam Study, 34 of the 54 SNPs were significantly associated with height at $p < 0.05$. Only for two SNPs the direction of (non-significant) height association was inconsistent with that reported by the original studies (Supplementary Table 4.3). Before estimating the potential of the genomic profile to predict human height, we have also tested if the 54 loci deviated from the within- or between-loci additivity assumption. After correction for multiple testing, we did not find statistically significant evidence for between-loci interactions (all nominal $p > 0.001$). Only one SNP (*rs*4794665 located in the NOG-RISK region) demonstrated significant deviation from a within-locus additive model after correction for multiple testing (corrected p = 0.0006, see Supplementary Table 4.3).

The genomic profile, based on 54 recently identified loci, was computed as the sum of the number of height-increasing alleles carried by a person, similar to Weedon et al. [5]. This profile explained 3.8% of the sex- and age-adjusted variation of height in the Rotterdam Study (Figure 3.2A). We also estimated the upper explanatory limit of the 54-loci allelic profile by defining the profile as a weighted sum of height-increasing alleles with weights proportional to the effects estimated in our own data using multivariable model (Supplementary Table 4.3). Such weighted genomic profile explained 5.6% of variation of height in the Rotterdam Study. The mean difference between people having "top" and "lowest" 5% of the genetic height score was 4.9 cm (Figure 3.2A, Table 3.1) [6.4 cm when using the weighted profile].

The ability of the genomic profile to predict a very tall (belonging to the upper 5% of the distribution) person was estimated using the Area Under the receiver-operating characteristic Curve (AUC) – a statistic routinely used to assess the predictive ability of a test in clinical practice [12, 13, 14, 15]. The AUC for the 54-loci genomic profile was 65% (68% for the weighted profile; Table 3.1 and Figure 3.3A).

Next, to estimate the predictive power of the Galtonian method, we used the family based Erasmus Rucphen Family (ERF) study [9], where parental height data were available for 550 participants. To construct the Galtonian predictive profile for every person for whom both paternal and maternal height was available, we computed the average of the parental height residuals. We

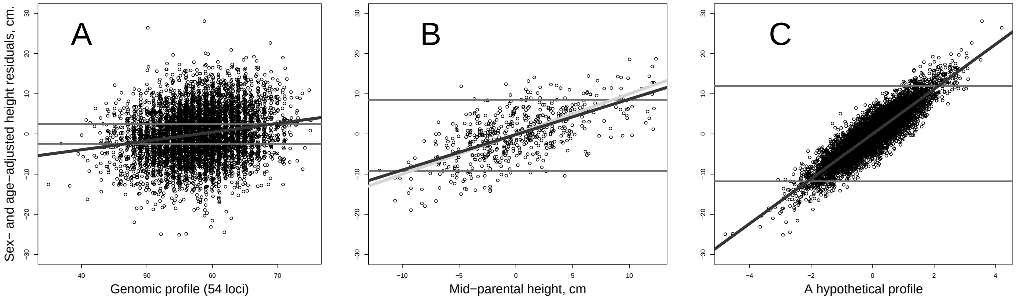


Figure 3.2: Observed sex- and age-adjusted height vs different predictive profiles. (a) Rotterdam Study, prediction with the genomic profile constructed from 54 loci, (b) ERF study, Galtonian prediction using mid-parental height values and (c) Rotterdam Study, a hypothetical profile explaining 80% of height variance. Red lines: mean residual height in people coming from top and bottom 5% of the profile distribution. Blue line: regression of the height residuals onto profile. In (b), green line has slope of 1, deviation of the blue line from the green showing "regression towards mediocrity".

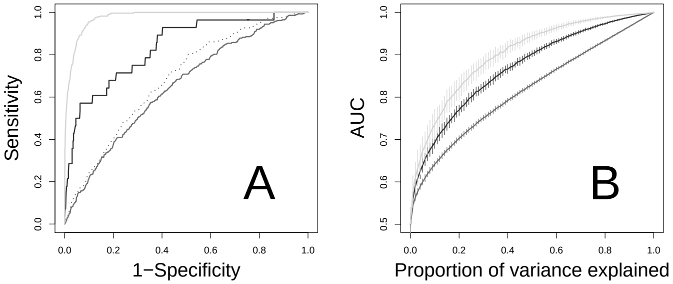


Figure 3.3: Accuracy to discriminate the top 5% tallest person, as measured by AUC, using different height profiles. (a) 54-loci genomic profile explaining 3.8% (54 loci, solid red line, AUC = 65% in the Rotterdam Study), population-specific 54-loci genomic profile explaining 5.8% in the Rotterdam Study (estimated using the data, red dotted line, $AUC = 68\%$ in the Rotterdam Study), mid-parental value explaining 40% (blue line, AUC=83% in the ERF study) and a hypothetical profile explaining 80% of height variance (green line, AUC97%). (b) AUC achieved by a test explaining certain proportion of height variance; red: predicting top 50%, blue: predicting top 5%, green: predicting top 1%. Vertical lines: standard error of the mean.

found that the proportion of height explained by Galtonian mid-parental profile was 40% (Figure 3.2B) - which is an order of magnitude higher than the result achieved using the 54-loci genomic profile. The mean difference between people having "top" and "lowest" 5% of the mid-parental predictive profile reached impressive 17.68 cm (Figure 3.2B, Table 3.1). Moreover, Galtonian prediction performed much better when discriminating very tall people (AUC=84%; Table 3.1 and Figure 3.3A). We have addressed the question if combing the parental height information with genotypic profile leads to better prediction. The analysis was restricted to 270 members of the ERF study for whom both parental phenotype and genetic data were available. Both mid-parental value ($p = 10^{-42}$) and the not weighted genomic profile ($p = 0.01$) were significantly associated with height of an offspring. Not surprisingly, the genomic profile was strongly correlated (Pearson's $\rho = 0.22$, $p = 0.0003$) with the mid-parental height value. Table 3.1 shows that while being statistically significant, considering the genomic profile added little to the prediction based on mid-parental values only (proportion of variance explained increased by $\sim 1.3\%$, and AUCs stayed virtually the same).

Finally we addressed the question of how much variance should a genomic profile explain to achieve a certain AUC value [15]. For this, using the Rotterdam Study data we simulated profiles explaining different proportion of trait variance, and evaluated AUCs for these (Figure 3.3B). For every evaluated point, one hundred simulations was performed. The simulations have shown that when one aims to predict a person having extreme (1% highest/lowest) value, a predictive profile explaining as little as 17% of trait's variance is sufficient to achieve an AUC of 80% (which may generally be considered as good for the screening purposes), and a profile explaining 53% to achieve an excellent AUC of 95%. On the other hand, good prediction of a person from the top/lower 5% trait's distribution requires a profile explaining 25% and excellent prediction of such person requires a profile explaining already 68% (Figure 3.3B).

It can be expected that if all loci controlling human height would be known a genomic profile could explain up to 80% of height variance. Under this scenario, the mean difference between people having "top" and "lowest" 5% of such hypothetic profile was $23.38 + 0.005$ cm (typical realization is presented in Figure 3.2C). As expected from the high proportion of explained variance, discriminative accuracy of this hypothetic profile was very high ($AUC = 97.4 + 0.3\%$, Table 3.1 and Figure 3.3A).

Discussion

In this work, we compared genomic and Victorian approaches to predict human height. In our data, the 54-loci genomic profile explained 4-6% and the Victorian Galton's mid-parental values explained 40% of the height variance. Adding genomic information to the mid-parental provided only small (1.3%) increase in proportion of variance explained. In forensics and human medicine, the question of binary classification of a person (e.g. "very tall" or not) based on some profile (score) is of large interest. We have previously proposed that the usefulness of a genomic profile associated with a binary outcome should be evaluated by the area under the receiver-operating characteristic (ROC) curve [14, 15]. In medicine, ROC analysis has been extensively used in evaluation of diagnostic tests. We show that 54-loci genomic profile had relatively low discriminative accuracy (AUC=65% for a person falling into 5% tallest). This value, however, is promising, e.g. approximately the same AUC is reached when predicting the risk of coronary heart disease using low density lipid levels [16]. We estimate that to achieve AUC of 80% using height genomic profiling, we need to explain at least three times the amount of variance currently explained with the available 54 loci. At the same time, the cheap and straightforward Galtonian approach demonstrated AUC of 84% when predicting 5% tallest person. The latter discriminative accuracy of 84% is better than that of many tests used in the clinical context, such as the Framingham risk scores that predicts coronary heart disease based on traditional risk factors such as blood pressure, lipid levels and smoking status [16]. However, the Galtonian prediction requires knowledge of parental height, which is not always available in applications such as e.g. forensics.

Our height study provides a strong example of a trait for which at the current stage a simple prediction based on phenotype of relatives clearly outperforms sophisticated genomic prediction. Would this hold for other phenotypes? The proportion of offspring's phenotypic variation which can be explained by mid-parental phenotypic value is $(h^2)^2/2$, where h^2 is heritability of the trait [17]. In a recent study we have estimated that 11 SNPs explain 3 to 5% of the variance of total cholesterol, and similar figures were obtained for high and low density lipoprotein cholesterol and triglicerides [18]. These traits typically exhibit about 30% heritability. Therefore Galtonian prediction cannot explain more then 5% of the trait's variance. Thus for lipid levels genomic prediction is already doing as good as (or as bad as) the Galtonian one. However, the genomic profiling, unlike the Galtonian, still has the potential to improve, as more loci affecting the phenotype of interest will be discovered.

While the upper limit for genomic profiling is determined by heritability, the genetic architecture of the trait is very important factor to consider in estimating the potential of predictive testing [15]. For example, for iris color single major locus explains the vast proportion of variance and the AUC of 80% is reached when predicting blue or brown iris color using only 3 SNPs [19]. On the other hand, for traits such as blood pressure only few loci explaining very small proportion of variance each are known and for such traits the prospects of genomic profiling are much worse. It can be expected that once all loci involved in human height will be revealed the discriminative accuracy of the genomic approach may surpass that of the Galtonian approach. However, it will be a tall order to find all these variants, at least using the current methodology consisting of (meta-analyses) of genome-wide association studies tailored to capture common variants. The 54 common variants discovered by now probably already include those with the largest effect sizes. Simply because of the fact that the variants with the larger effect sizes are most easily captured, the detection of new height genes will require progressively bigger sample sizes (e.g. to detect a locus explaining 0.1% of variance at genome-wide significant $p < 5 \cdot 10^{-8}$ with power of 80%, one would need to study $40,000$ people, while to detect a locus explaining 0.01% one would need $400,000$ people) [5].

As noted by Galton, "stature is not a simple element, but a sum of accumulated lengths and thicknesses of more than a hundred of bodily parts" The beautiful regularity in the statures of a population " is due to the number of variable elements the stature is the sum [1]. Detailed analysis of the factors controlling these endophenotypes is likely to be necessary to discover new loci and make genetic findings useful for applications in forensics and medicine.

We conclude that while the genomic approach is potentially more powerful than the Victorian Galton's method, the latter will long stay unsurpassed both in terms of discriminative accuracy and costs when the trait in question is highly heritable and the parental phenotype is usually available. For less heritable traits, such as lipid levels, and in situations when parental information is not available (e.g. forensics), genomic methods may provide an alternative, given the variants determining an essential proportion of variation could be identified.

Acknowledgements

We thank Prof. T. Frayling for useful discussion and comments and Dr. M. Bevova for comments and graphical art support. We acknowledge the work of Dr. F. Rivadeneira, Dr. M. Moorhouse, P. Arp and M. Jhamai, who have created the Rotterdam Study genotypic database. This research was supported by grants from the Netherlands Foundation for Scientific Research (NWO), a joint grant from NWO and the Russian Foundation for Basic Research (RFBR), by the Centre for Medical Systems Biology (CMSB) and by the Netherlands Forensic Institute (NFI). MNW is a Vandervell Foundation Research Fellow. Genome-wide genotyping of Rotterdam Study is supported by NWO (175.010.2005.011).

BIBLIOGRAPHY

1. F, G. Regression towards mediocrity in hereditary stature. *Journal of the anthropological institute* **15,** 246–263 (1886).
2. RA, F. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* **52,** 399–433 (1918).
3. PM, V., SE, M. & MA, F. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* **2,** e41 (2006).
4. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 2001).
5. MN, W., H, L. & CM, L. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40,** 575–583 (2008).
6. G, L., AU, J. & C, G. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40,** 584–591 (2008).
7. DF, G., GB, W. & G, T. Many sequence variants affecting diversity of adult human height. *Nat Genet* **40,** 609–615 (2008).
8. A, H., MM, B. & van Duijn CM. The Rotterdam Study: objectives and design update. *Eur J Epidemiol* **22,** 819–829 (2007).
9. LM, P., I, M., B, O., van Duijn CM & YS, A. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69,** 288–295 (2005).
10. YS, A., P, H. & I, M. Linkage disequilibrium in young genetically isolated Dutch population. *Eur J Hum Genet* **12,** 527–534 (2004).
11. Y, L. & GR, A. Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet* **S79,** 2290 (2006).
12. JA, H. & BJ, M. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (1982).
13. A, A. & M, S. R2: a useful measure of model performance when predicting a dichotomous outcome. *Stat Med* **18,** 375–384 (1999).
14. AC, J., MC, P., EW, S. & van Duijn CM. Revisiting the clinical validity of multiplex genetic testing in complex diseases. *Am J Hum Genet* **74,** 585–588 (2004).
15. AC, J. *et al.* Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* **8,** 395–400 (2006).

16. PW, W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97,** 1837–1847 (1998).

17. DS, F. & TFC, M. 1996.

18. Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41,** 47–55 (Jan. 2009).

19. M, K., F, L. & AC, J. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* **82,** 411–423 (2008).

3.2 THE ROLE OF COMMON GENETIC AND ENVIRONMENTAL FACTORS IN EXTREME HIGH AND LOW LEVELS OF TOTAL CHOLESTEROL

Struchalin M, Lennart C Karssen, Najaf Amin, Kelly S Benke, Abbas Dehghan, Jacqueline C Witteman, Albert Hofman, Ben A Oostra, Oscar H Franco Duran, Cornelia M van Duijn

Department of Epidemiology, Erasmus MC, Postbus 2040, 3000 CA Rotterdam, The Netherlands

Abstract

Serum concentration of total cholesterol (TC) is a complex trait, for which genome-wide association studies discovered common genetic variants explaining about 30% of TC genetic variation. In view of personalized medicine, the most important persons to target in the population are those with extreme levels of TC (low and high) in the population. These are the persons to be treated aggressively. We determined the extent to which 52 known common genetic variants and environmental factors (i.e. body mass index, smoking, alcohol intake, diabetes, age and gender) explain phenotypic variation in extreme TC levels in a family based-study i.e the Erasmus Rucphen Family ($N = 2,239$) and a population-based study i.e. the Rotterdam Study ($N = 5,441$). We studied proportion of individuals of 5%, 10% and 20% from the top and the bottom extremes of the TC distribution. The discriminative power was evaluated by the Receiver Operating Curve (ROC) and corresponding Area Under the Curve (AUC). We observe no substantial decrease in TC variation explained by studied factors in the extreme TC levels. The current study implies GWAS as a primary method for discovering new TC associated genetic variants which will explain a substantial part of TC genetic variation.

In preparation

3.3 ASSOCIATIONS BETWEEN RECENTLY DISCOVERED GENETIC VARIATIONS IN METABOLIC TRAITS AND ARTERIAL STENOSIS IN PATIENTS WITH RECENT CEREBRAL ISCHEMIA

E.G. van den Herik[1,3], M. Struchalin[2,3], L.M.L. de Lau[1], H.M. den Hertog[1], S. Fonville[1], P.J. Koudstaal[1], C.M. van Duijn[2]

[1] *Departments of Neurology, Erasmus MC University Medical Center, Rotterdam, The Netherlands*
[2] *Departments of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands*
[3] *These authors contributed equally to the work.*

Abstract

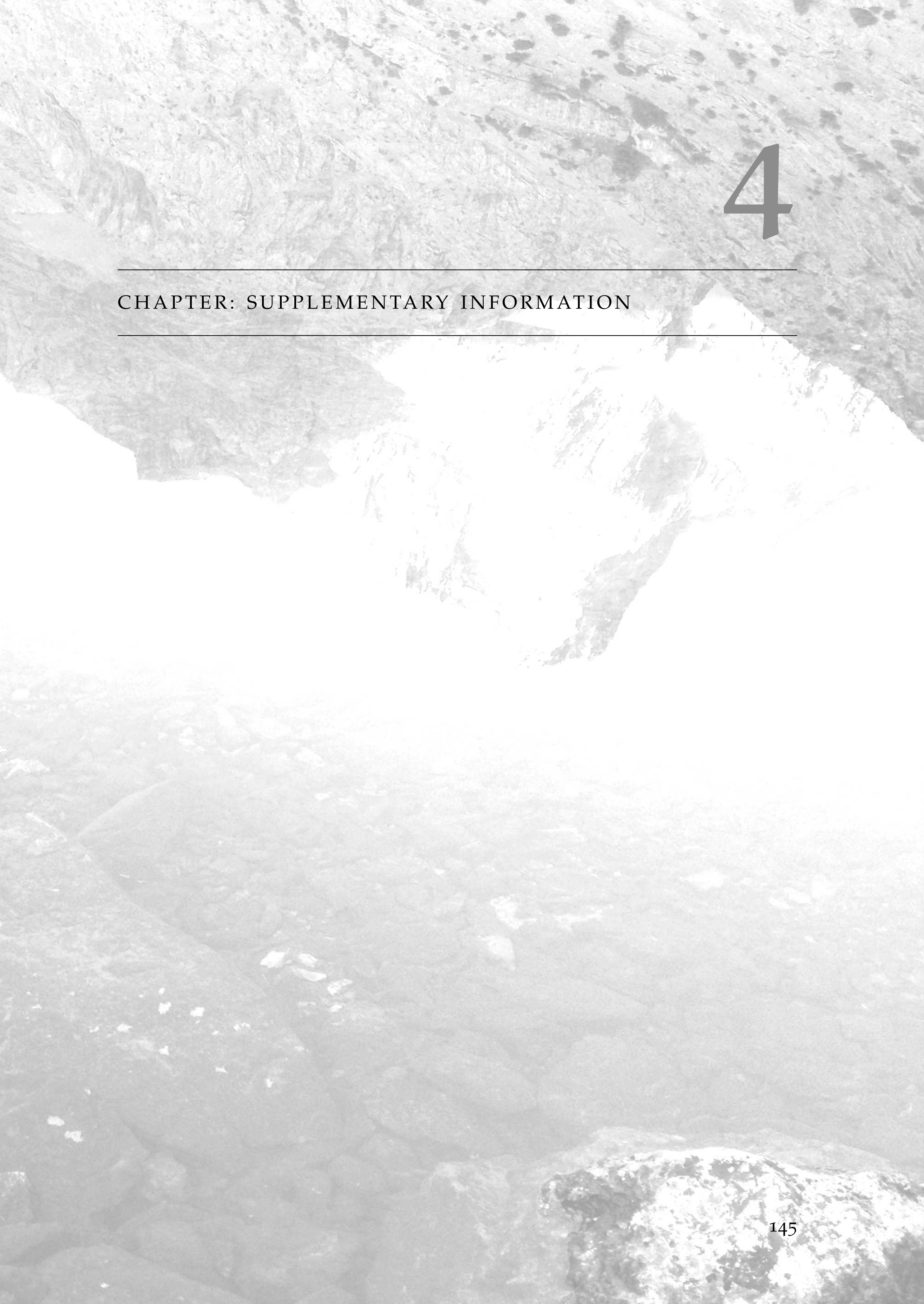
BACKGROUND: Recent large genome-wide association studies have found many new genes to be associated with metabolic traits including hypertension, lipid levels and diabetes. A next step after gene discovery is to investigate associations between these genes and clinically relevant endpoints. Therefore, we studied the relation between recently discovered genetic variations in metabolic traits and craniocervical artery stenosis, in patients with transient ischemic attack or ischemic stroke.

METHODS: We included 700 patients with a recent transient ischemic attack or ischemic stroke. In all patients CT-angiography from the aortic arch to the intracranial vessels was performed and scored for degree of stenosis in each artery. Our primary outcome was presence of a stenosis $\geq 30\%$ in any artery. Genotyping was performed with Metabochip, a targeted gene chip for metabolic traits containing $\approx 200,000$ SNPs.

RESULTS: Five loci were found to be strongly associated with presence of stenosis (*GLIS3*, $p = 1.6 \cdot 10^{-6}$; *AGBL2*, $p = 1.2 \cdot 10^{-4}$; *SBF2*, $p = 7.9 \cdot 10^{-5}$; *SCAMP5*, $p = 8.7 \cdot 10^{-5}$; and *MC4R*, $p = 1.7 \cdot 10^{-4}$). After correction for multiple testing using permutations only the *GLIS3* locus remained. For this gene, the risk allele was associated with a 2.2 (95%CI $1.6 - 3.1$)-fold increase of stenosis risk. This gene was previously shown to lead to elevated fasting glucose levels. We found no significant association between loci implicated in hypertension or hypercholesterolemia and presence of stenosis.

CONCLUSIONS: Our study suggests an association between atherosclerosis and a gene implicated in increased fasting glucose levels, but not with blood pressure- or lipid-related genes, in patients with recent cerebral ischemia.

In preparation



4

CHAPTER: SUPPLEMENTARY INFORMATION

Text S1. Laboratory details for `MC1R` SNP genotyping.

After optimization, 2 ng of dried DNA in clear 384-well plates (Applied Biosystems, Foster City, USA) was typed using a reaction volume of $5\mu l$. The reaction for SNP rs1805007 contains $1x$ LightCycler 480 genotyping master (Roche, Mannheim, Germany), $0.5\mu M$ forward primer, $1.0\mu M$ reverse primer, $0.2\mu M$ $3' - FL$ labeled probe and $0.2\mu M$ $5' - LC$ labeled probe. PCR was performed in a Lightcycler 480 (Roche) at 95°C for 10 minutes followed by 45 cycles amplification of 95°C for 10 seconds, 57°C for 10 seconds, 72°C for 15 seconds, followed by the melting curve from 40°C to 80°C. Because SNP rs1805008 proved to be problematic, due to many strong binding stem loops present in the area, Tib Molbiol designed an internal labeled primer which excludes some loops and breaks up the strongest inside of the amplification. This primer also functions as the sensor probe in the reaction. The reaction for SNP rs1805008 contains $1x$ LightCycler 480 genotyping master (Roche), $0.25\mu M$ forward primer, $0.5\mu M$reverse primer and $0.2\mu M$ $36' - FL$ labeled probe. The amplification was performed in a Lightcycler 480 (Roche) at 95° C for 10 minutes followed by 50 cycles amplification of 95° C for 10 seconds, 55° C for 10 seconds, 72° C for 10 seconds, followed by the melting curve from 40° C to 80° C.

Figures and tables

Table 4.1: **Primers of two `MC1R` SNPs.**

Primer/probe	rs1805007	rs1805008
Forward primer	TGTCGGACCTGCTGGTGAG	CTCCATGCTGTCCAGCCTC
Reverse primer	ACGTGGTCGTAGTAGGCGATGA	CGCAACGGC XTCGACGC (internal labeled)
5′-LC Labeled probe	640-CACTGCGCTACCACAGC p	
3′-FL Labeled probe	TGGACCGCTACATCTCCATCTTCTAC-FL	GTGACCCTGCCGCGGGC-FL

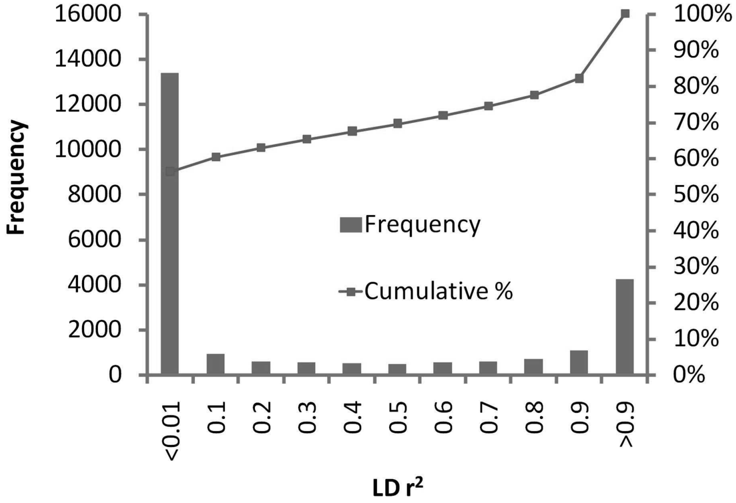


Figure 4.1: **The LD r^2 distribution of the physically close and rare SNP pairs on Illumina 550K chip.**

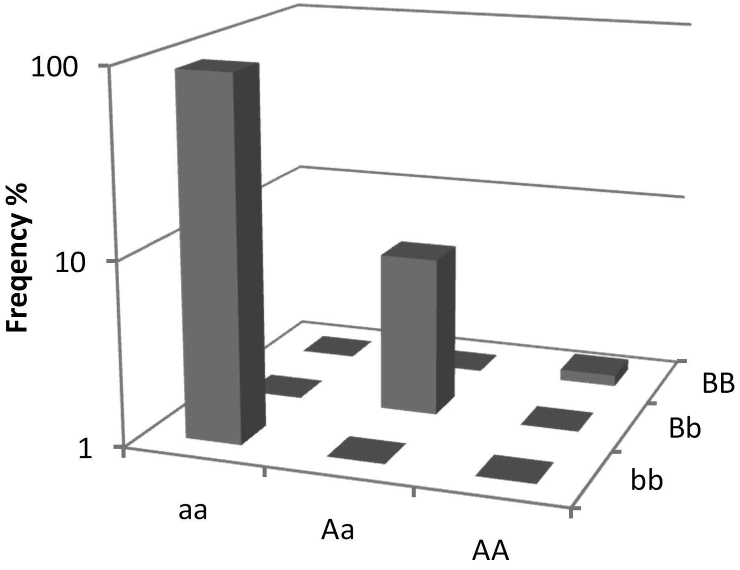


Figure 4.2: **Cross-genotypes between 2 rare SNPs in high LD.**

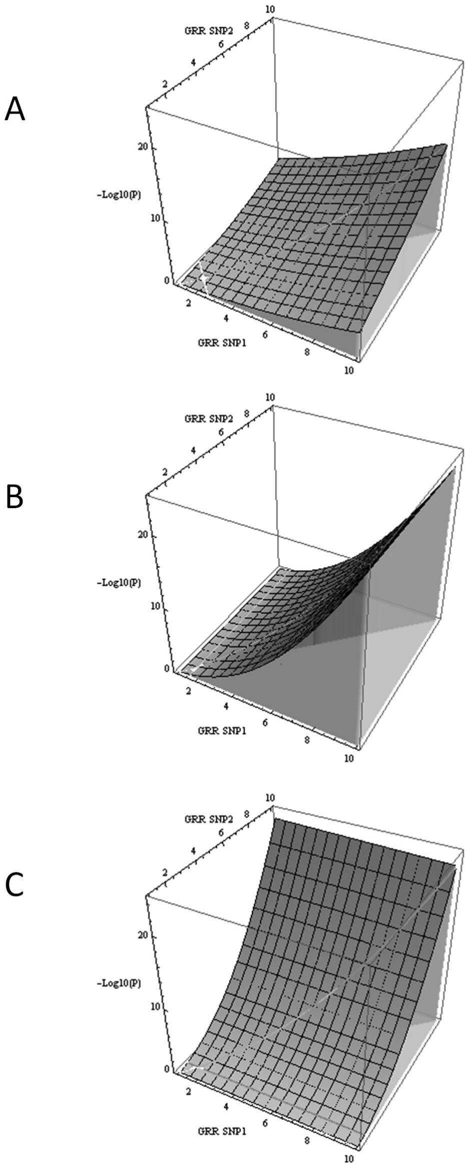
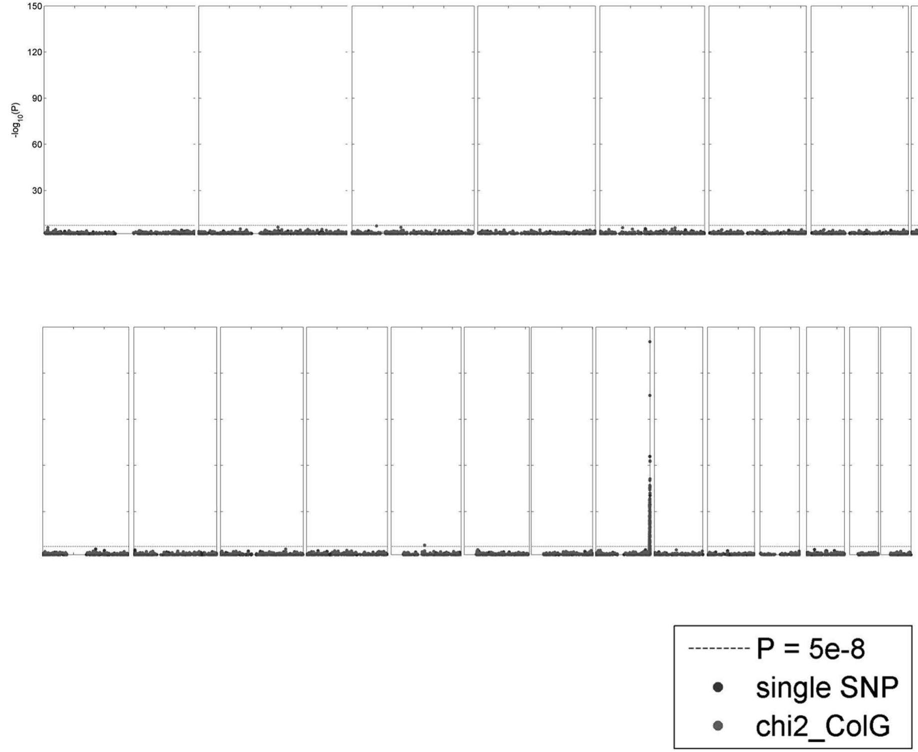


Figure 4.3: **Expected P-values from CDH and single SNP analyses considering 2 recessive SNPs independently associated with phenotype.** The $-\log_{10}(P)$ values for CDH test (A) and single SNP analysis (B and C) are plotted against the genotype relative risks of homozygote causal allele (GRR ranging from 1 to 10). Other parameters are fixed (the frequencies of causal alleles $= 0.05$, $N = 10,000$, $\alpha = 5\%$).

Figure 4.4: **Manhattan plot showing association with the red-hair color phenotype in the Rotterdam Study.** The $-\log_{10}(P)$ values for association with red hair color are plotted for each genotyped SNP according to its chromosomal position (blue dots) and for the CDH test in each sliding window consisting of 100 SNPs (green dots).



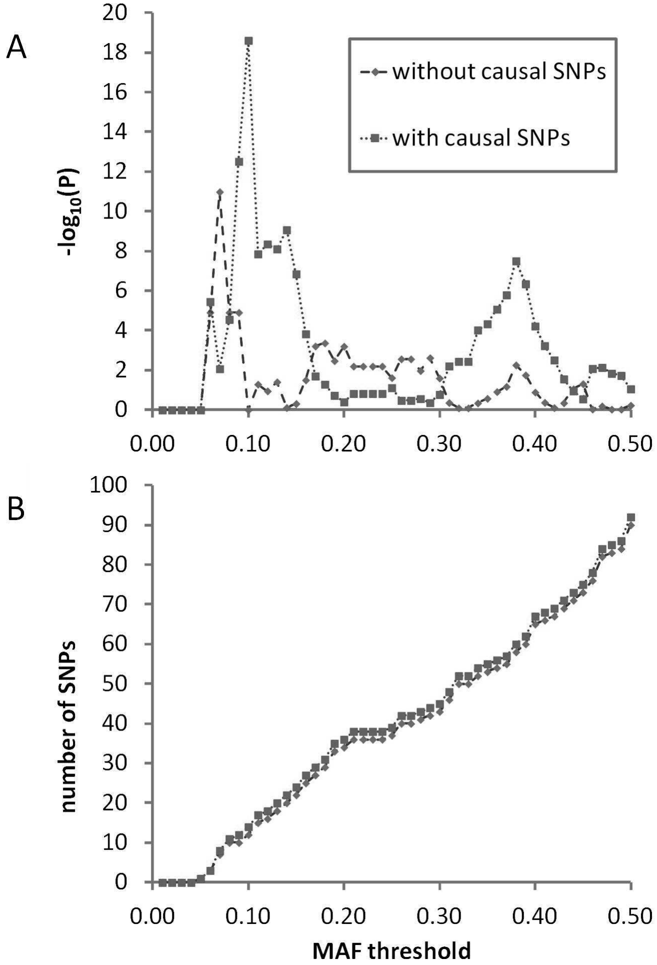
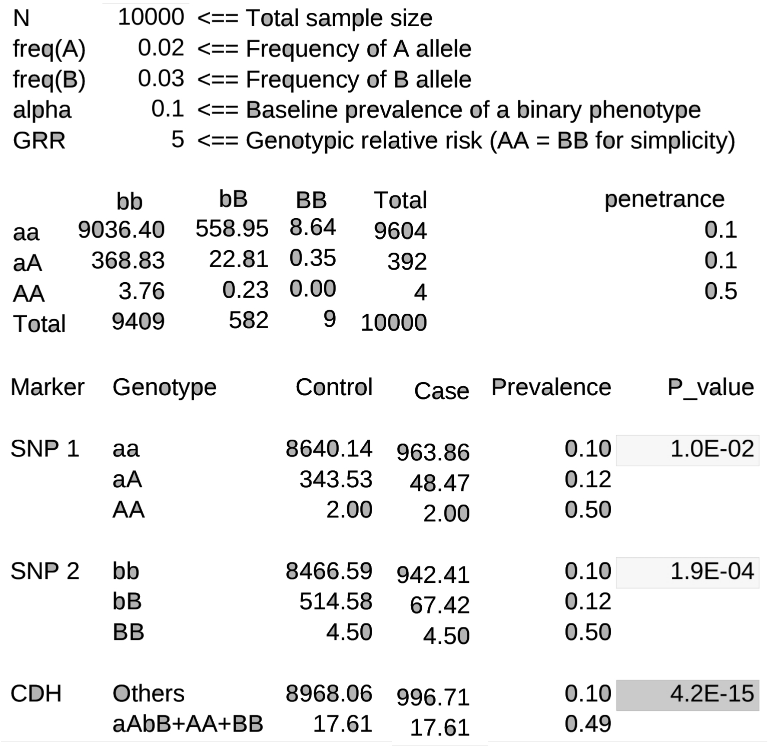


Figure 4.5: **Association analysis of the `MC1R` SNPs and the red hair color using the weighted sum statistic (WSS).** All number of genotyped SNPs in the 87.88 to 88.69 Mb region of (N SNPs = 90) were included to the WSS analysis according to the minor allele frequencies in the ascending order. The $-\log_{10}(P)$ values from WSS were plotted against the MAF thresholds (blue dots). The analysis was then repeated by assuming that two causal SNPs rs1805007 and rs1805008 were available on the chip (red dots). A, the $-\log_{10}(P)$ values; B, the number of SNPs included in the analysis.

Table 4.2: **The expected *P*-values from CDH and single SNP analysis.**



N	10000	<== Total sample size
freq(A)	0.02	<== Frequency of A allele
freq(B)	0.03	<== Frequency of B allele
alpha	0.1	<== Baseline prevalence of a binary phenotype
GRR	5	<== Genotypic relative risk (AA = BB for simplicity)

	bb	bB	BB	Total	penetrance
aa	9036.40	558.95	8.64	9604	0.1
aA	368.83	22.81	0.35	392	0.1
AA	3.76	0.23	0.00	4	0.5
Total	9409	582	9	10000	

Marker	Genotype	Control	Case	Prevalence	P_value
SNP 1	aa	8640.14	963.86	0.10	1.0E-02
	aA	343.53	48.47	0.12	
	AA	2.00	2.00	0.50	
SNP 2	bb	8466.59	942.41	0.10	1.9E-04
	bB	514.58	67.42	0.12	
	BB	4.50	4.50	0.50	
CDH	Others	8968.06	996.71	0.10	4.2E-15
	aAbB+AA+BB	17.61	17.61	0.49	

It was shown in the results of type I error of investigated variance homogeneity tests that in a case of SNP effect presence rank transformation to normality of a trait which follows non-normal distribution results to perfectly normally distributed trait whereas distribution of each genotypic groups becomes distorted. Figure S4.6 explains such a deformation.

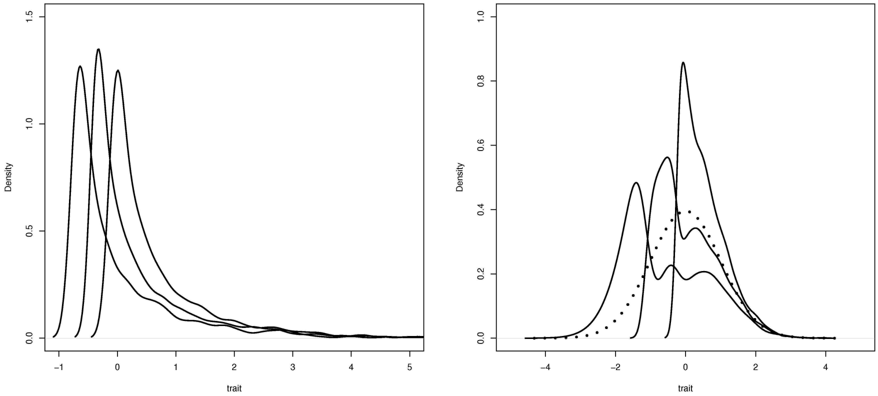


Figure 4.6: **Distribution of a trait for each genotypic groups (bold curves) and for all groups together (dotted curve) before transformation to normality of a trait (left) and after transformation (right).**

Analytical expressions for variances of trait's distribution in each genotypic group were obtained in this work. They can be used to obtain dependence of non-centrality parameter (and therefore power) on model parameters. To validate these analytical results simulations were done. Figure S4.7 shows analytical curves and simulated points for dependence of power on interaction effect for direct test and variance homogeneity tests.

Figures S4.8, S4.9, and S4.10 shows dependence of non-centrality parameter of variance homogeneity test on effect of a factor for a cases of one degree of freedom tests: when AA is tested against AB and BB, AB against AA and BB, BB against AA and AB.

Figures S4.11 show dependence of power of variance homogeneity test with two degrees of freedom on interaction effect for threshold α corresponding to $5 \cdot 10^{-8}$ and 0.01.

Figures S4.12 show Genome-wide $-\log 10(P - \text{value})$ and Q-Q plot for Levene's variance homogeneity test applied for the Rotterdam Study. Q-Q plot presents

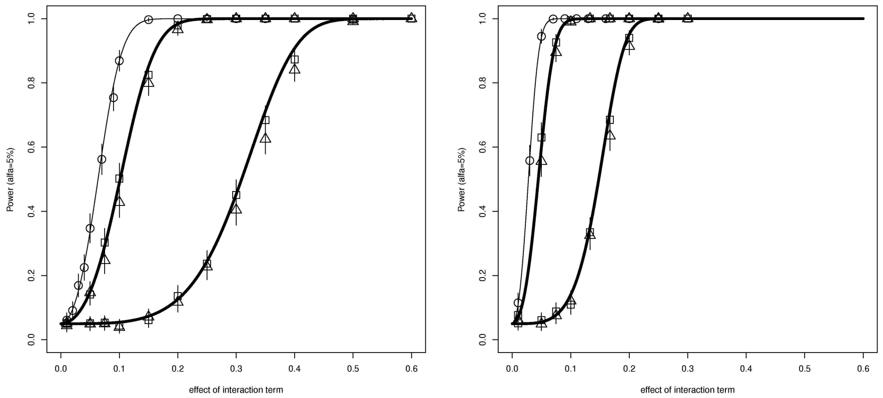


Figure 4.7: **Dependence of power on interaction effect for direct test and different variance homogeneity tests.** The thin curves on the left in each subplot corresponds to analytical expression of power of direct test. The two bold curves in each subplot corresponds to analytical expression of variance homogeneity test. The left bold curve is for a case of absence effect of a factor, the right bold curve is for effect of a factor one. The points correspond to simulations for direct (circles), Bartlett's (squares) and Levene's (triangles) tests. The left plot is for the case when frequency of interacting allele is 5% and no SNP effect. The right plot is for interacting allele frequency 50% and snp effect 0.3.

only those SNPs which have three genotypes. As one can see there is no SNPs reached genome-wide significance level.

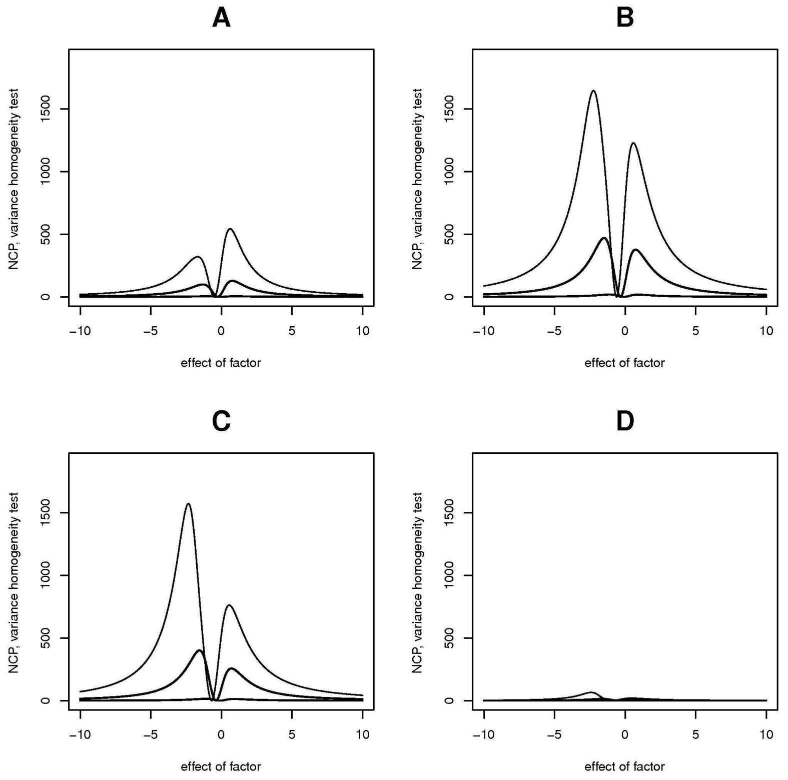


Figure 4.8: **Dependence of non-centrality parameter of variance homogeneity test on effect of a factor for a case when group AA is tested against AB and BB.** The top curve on each plot shows results for interaction effect equals $\beta_{gF} = 1$, the middle curve is for $\beta_{gF} = 0.5$, and the bottom curve is for $\beta_{gF} = 0.1$. Each subplot shows different frequency of interacting allele. (A – 0.05, B – 0.4, C – 0.6, D – 0.95).

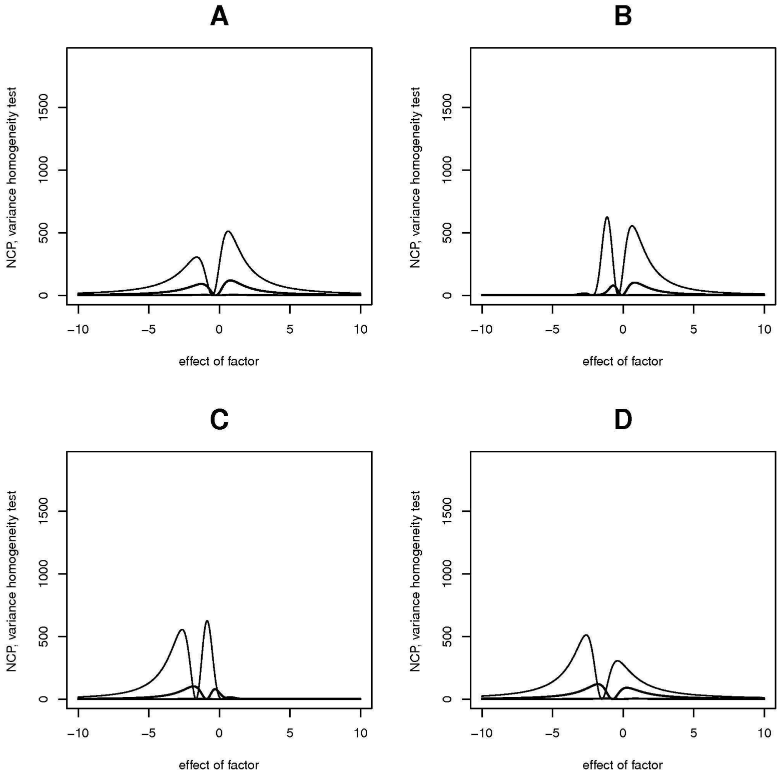


Figure 4.9: **Dependence of non-centrality parameter of variance homogeneity test on effect of a factor for a case when group AB is tested against AA and BB.** The top curve on each plot shows results for interaction effect equals $\beta_{gF} = 1$, the middle curve is for $\beta_{gF} = 0.5$, and the bottom curve is for $\beta_{gF} = 0.1$. Each subplot shows different frequency of interacting allele. (A – 0.05, B – 0.4, C – 0.6, D – 0.95).

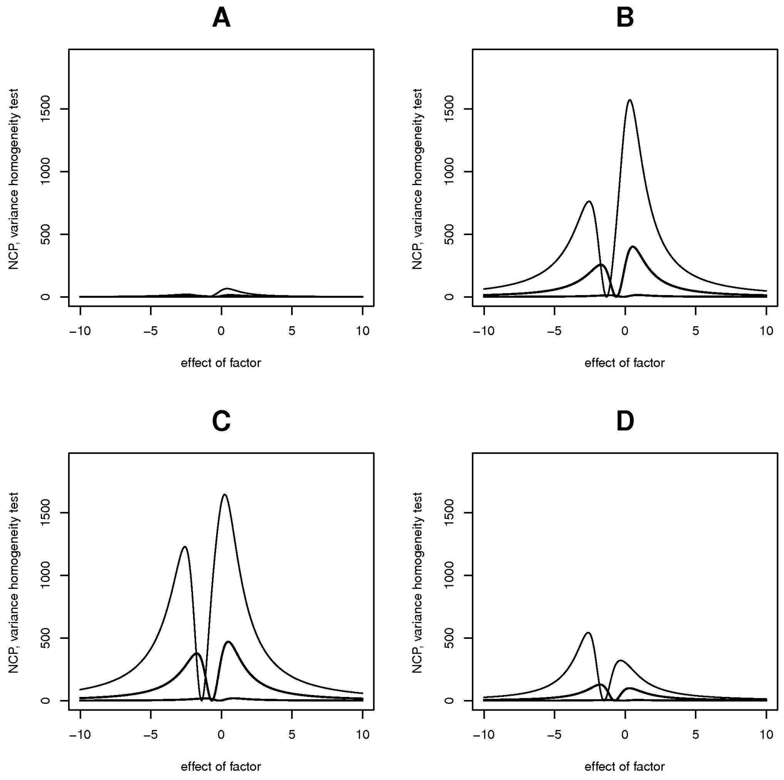


Figure 4.10: **Dependence of non-centrality parameter of variance homogeneity test on effect of a factor for a case when group BB is tested against AA and AB.** The top curve on each plot shows results for interaction effect equals $\beta_{gF} = 1$, the middle curve is for $\beta_{gF} = 0.5$, and the bottom curve is for $\beta_{gF} = 0.1$. Each subplot shows different frequency of interacting allele. (A – 0.05, B – 0.4, C – 0.6, D – 0.95).

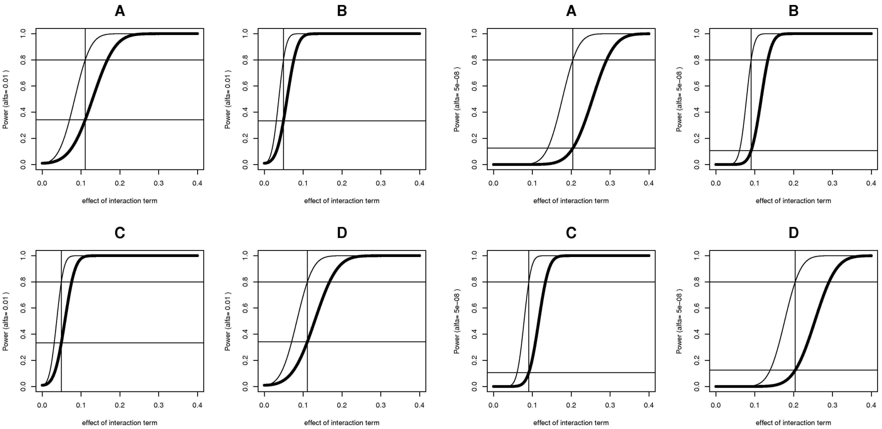


Figure 4.11: **Dependence of power of variance homogeneity test on interaction effect for threshold α corresponding to $5 \cdot 10^{-8}$ (left four plots) and 0.01 (right four plots).** Thin curve in each subplot corresponds to direct test, bold curve corresponds to upper limit of variance homogeneity test. Each subplot shows different frequency of interacting allele (A – 5%, B – 40%, C – 60%, D – 95%)

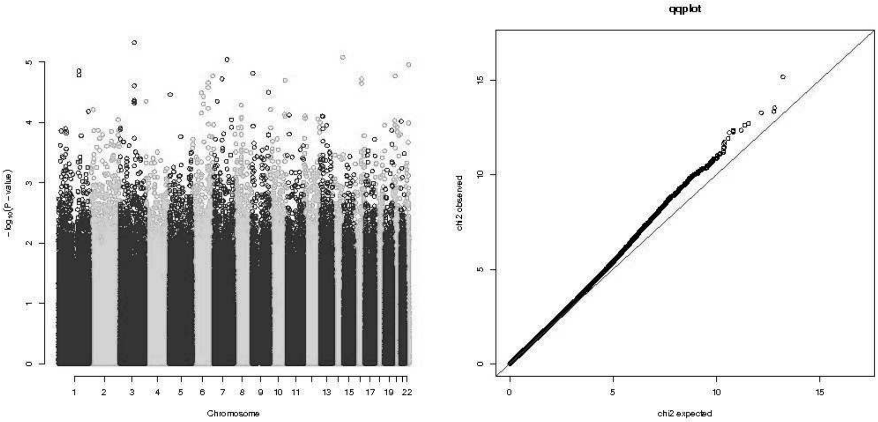


Figure 4.12: **Genome-wide log(p-value) and Q-Q plot for Levene's variance homogeneity test applied for the Rotterdam Study.**

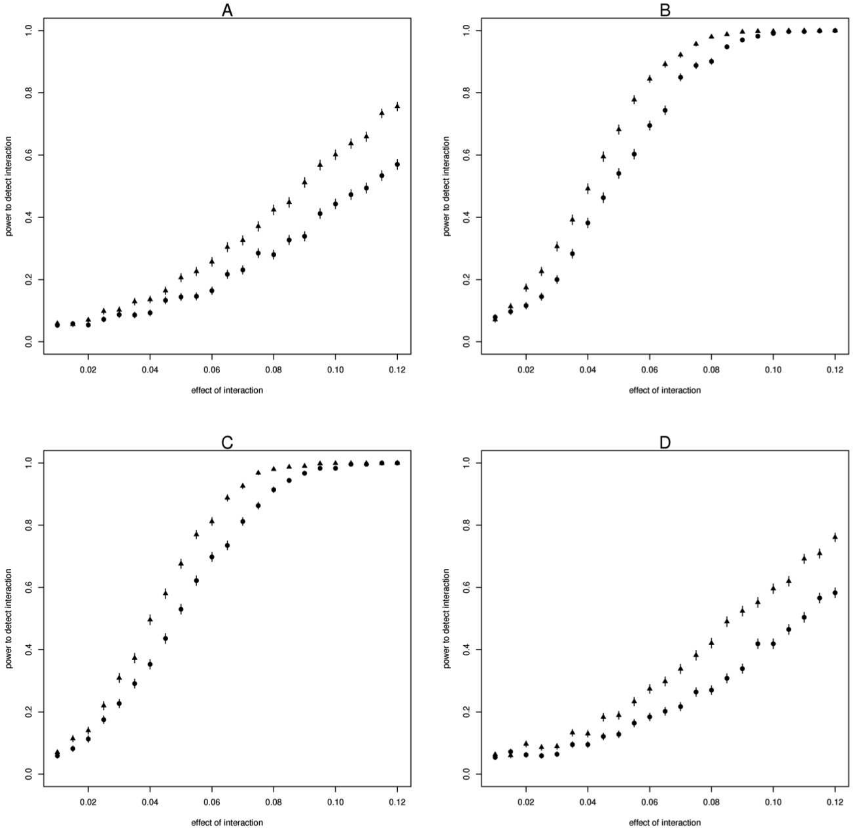


Figure 4.13: **Power to detect variance heterogeneity induced by interaction, assuming Normal distribution of residual error.** Dependency of power to detect variance heterogeneity induced by interaction and the effect of interaction, β_{gF} , using Levene's (circles) and SVLM (triangles) tests. The residual error follows Normal distribution. Scenarios with different frequencies of interacting allele are given in Panel A – 5%, Panel B – 40%, Panel C – 60%, and Panel D – 95%.

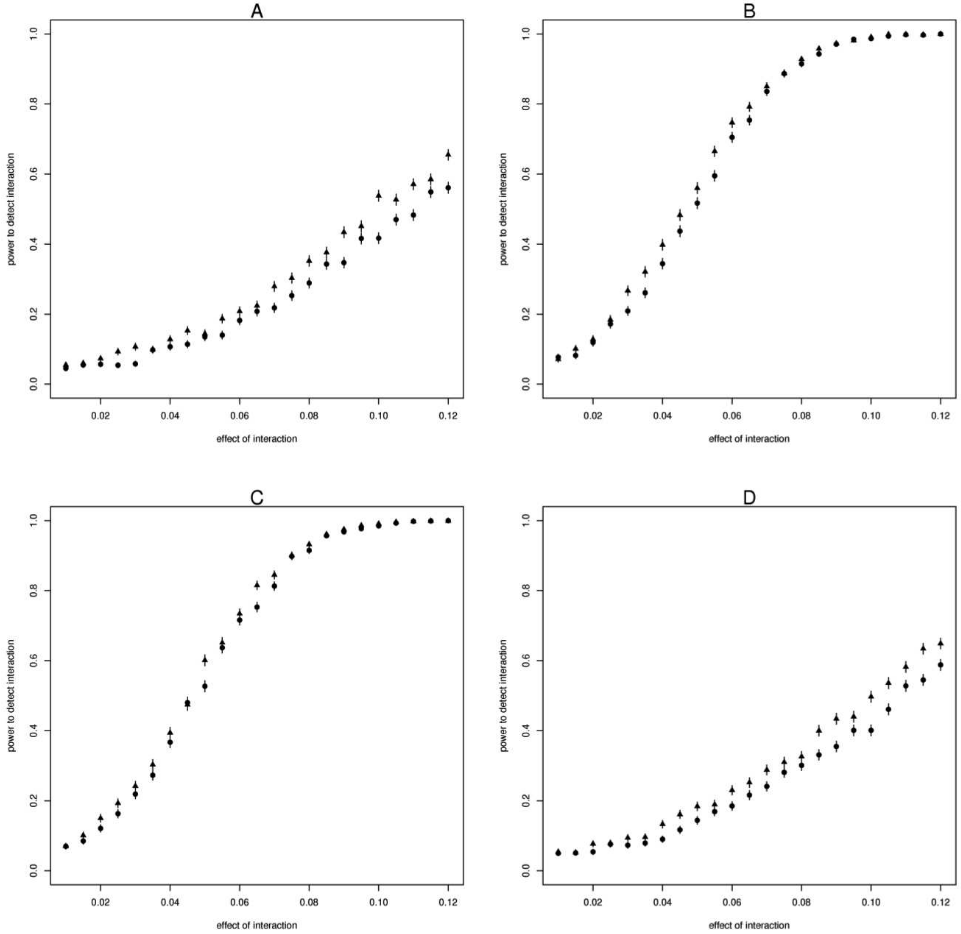


Figure 4.14: **Power to detect variance heterogeneity induced by interaction, assuming $\chi^2_{df=5}$ distribution of residual error.** Dependency of power to detect variance heterogeneity induced by interaction and the effect of interaction, β_{gF} , using Levene's (circles) and SVLM (triangles) tests. The residual error follows Normal distribution. Scenarios with different frequencies of interacting allele are given in Panel A – 5%, Panel B – 40%, Panel C – 60%, and Panel D – 95%.

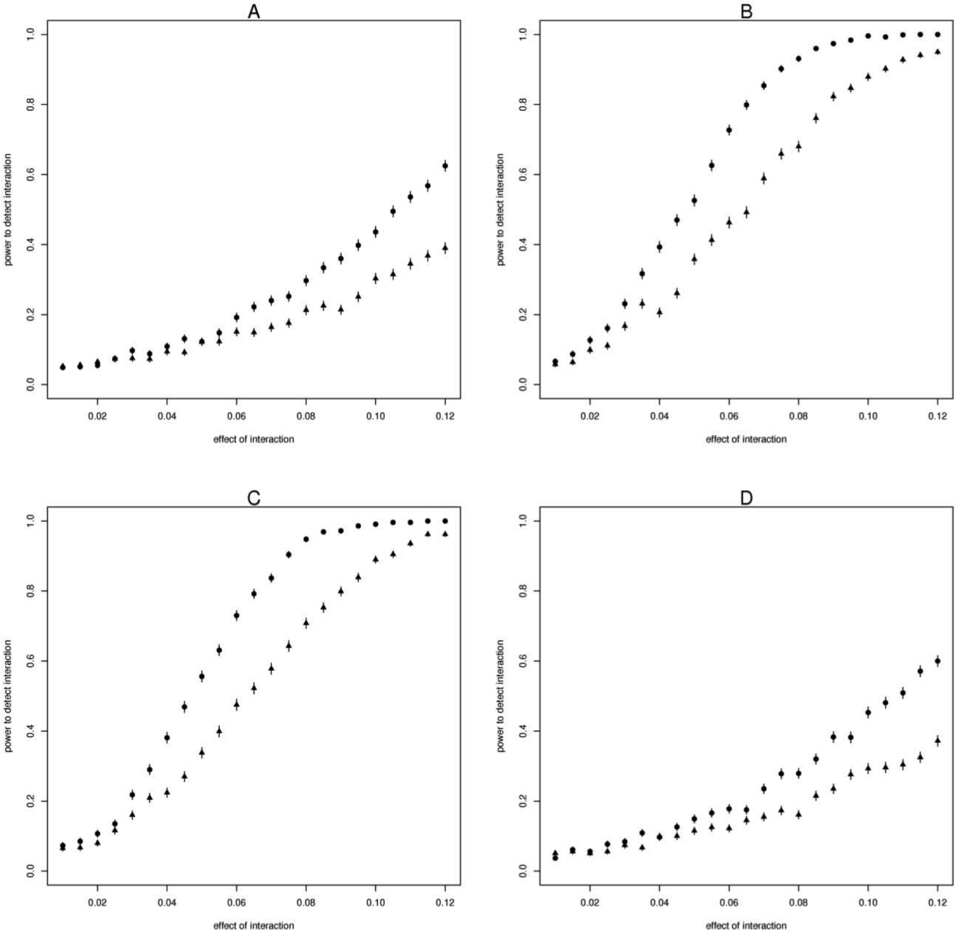


Figure 4.15: **Power to detect variance heterogeneity induced by interaction, assuming $\chi^2_{df=1}$ distribution of residual error.** Dependency of power to detect variance heterogeneity induced by interaction and the effect of interaction, β_{gF} , using Levene's (circles) and SVLM (triangles) tests. The residual error follows Normal distribution. Scenarios with different frequencies of interacting allele are given in Panel A – 5%, Panel B – 40%, Panel C – 60%, and Panel D – 95%.

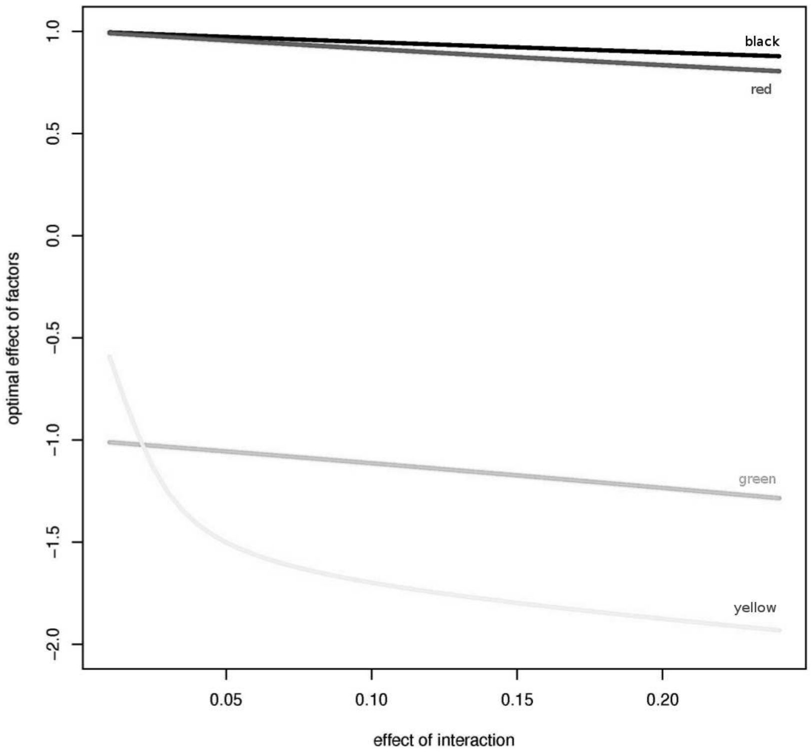


Figure 4.16: **Optimal effect of the factor F (β_F) as a function of the interaction effect (β_{gF})** The value of optimal effect of interacting factor F, β_F , as a function of the effect of interaction, β_{gF} for allele frequencies 5% (black), 40% (red), 60% (green) and 95% (yellow).

Table 4.3

									Single SNP association				Joint estimation			Single SNP association,					
					Original		RS		additive model*			additive model**			general model***					
SNP	Allele+	Allele-	gene	N	P	Afreq+	Typed	R2	Beta+	se+	p	Beta	se	p	B	beta_AB	se_AB	beta_BB	se_BB	P1****
rs6440003	A	G	ZBTB38	30147	1.8E-24	0.485	no	0.990	0.494	0.117	0.000	9.274	7.667	0.226	G	-0.607	0.207	-0.994	0.234	0.506
rs2282978	C	T	CDK6,PEX1,GATAD1,ERVWE1	30147	7.8E-23	0.343	yes	1.000	0.466	0.122	0.000	0.253	0.160	0.115	T	-0.293	0.267	-0.849	0.268	0.468
rs16896068	G	A	LCORL	30147	2.4E-13	0.856	no	0.999	0.486	0.167	0.004	-3.297	11.209	0.769	A	-0.530	0.190	-0.719	0.611	0.622
rs4549631	C	T	LOC387103	30147	4.7E-13	0.494	no	1.000	0.214	0.116	0.066	0.186	0.114	0.103	C	0.316	0.200	0.425	0.233	0.526
rs3791675	C	T	EFEMP1	30147	2.2E-12	0.761	no	0.994	0.222	0.137	0.105	5.443	2.314	0.019	C	0.355	0.374	0.535	0.364	0.702
rs2814993	A	G	C6orf106	30147	4.1E-12	0.155	yes	1.000	0.632	0.159	0.000	0.517	0.159	0.001	G	-1.240	0.543	-1.757	0.528	0.242
rs10512248	G	T	PTCH1	30147	4.2E-11	0.332	no	0.908	0.263	0.129	0.042	0.248	0.127	0.050	T	-0.277	0.302	-0.533	0.293	0.959
rs12735613	G	A	SPAG17	30147	4.4E-11	0.798	no	0.999	0.279	0.145	0.054	0.312	0.142	0.028	A	-0.236	0.177	-0.689	0.423	0.670
rs11107116	T	G	SOCS2	30147	5.6E-10	0.232	yes	1.000	0.296	0.136	0.029	0.248	0.133	0.062	G	-0.162	0.367	-0.502	0.356	0.696
rs1390401	A	G	ZNF678	30147	5.4E-09	0.810	no	1.000	0.102	0.148	0.490	0.104	0.146	0.475	A	0.227	0.462	0.301	0.449	0.775
rs3116602	T	G	DLEU7	30147	6.8E-09	0.818	no	0.932	0.726	0.154	0.000	0.728	0.152	0.000	T	1.224	0.496	1.834	0.476	0.291
rs6686842	T	C	SCMH1	30147	1.7E-08	0.435	no	1.000	0.299	0.117	0.010	0.299	0.114	0.009	T	0.234	0.186	0.618	0.238	0.652
rs10906982	A	T	ADAMTSL3	30147	1.7E-08	0.521	no	0.999	0.335	0.116	0.004	0.331	0.126	0.009	T	-0.362	0.195	-0.667	0.232	0.860
rs6724465	G	A	IHH	30147	2.1E-08	0.900	no	0.998	0.471	0.191	0.014	0.454	0.187	0.015	G	1.584	0.769	1.905	0.749	0.135
rs10935120	G	G	ANAPC13orCEP63	30147	7.3E-08	0.691	no	0.999	-0.046	0.125	0.711	-0.026	0.123	0.830	G	0.233	0.292	0.060	0.289	0.290
rs8041863	A	T	ACAN	30147	8.1E-08	0.456	no	0.987	0.111	0.118	0.345	0.088	0.115	0.446	T	-0.125	0.217	-0.224	0.237	0.941
rs8099594	A	G	DYM	30147	3.1E-07	0.663	no	0.945	0.286	0.125	0.023	0.279	0.123	0.023	A	0.272	0.284	0.564	0.279	0.957
rs11205277	G	A	HistoneClass2A,MTMR11,SV2A,SF3B4	36485	1.4E-10	0.450	yes	0.995	0.365	0.117	0.002	0.377	0.115	0.001	A	-0.308	0.217	-0.720	0.237	0.755
rs678962	G	T	DNM3	33992	3.2E-08	0.214	yes	1.000	0.130	0.143	0.365	0.132	0.141	0.348	G	-0.009	0.175	0.684	0.418	0.163
rs2274432	T	G	C1orf19,GLT25D2	36485	7.8E-09	0.365	no	0.998	0.442	0.119	0.000	0.442	0.117	0.000	T	0.455	0.177	0.876	0.255	0.923
rs3791679	T	G	EFEMP1,PNPT1	39509	5.9E-11	0.761	yes	0.999	0.199	0.137	0.146	-5.216	2.311	0.024	T	0.366	0.373	0.514	0.364	0.630
rs6763931	A	G	ZBTB38	39509	1.4E-27	0.485	yes	0.998	0.490	0.116	0.000	-37.565	31.257	0.229	G	-0.601	0.206	-0.987	0.233	0.513
rs6830062	T	C	LCORL,NCAPG	39509	1.3E-10	0.856	yes	0.999	0.486	0.167	0.004	3.781	11.209	0.736	C	-0.531	0.190	-0.721	0.611	0.623
rs1812175	C	A	HHIP	39509	9.7E-12	0.843	yes	1.000	0.258	0.159	0.104	0.047	0.171	0.782	C	1.012	0.533	1.122	0.517	0.138
rs12198986	A	G	BMP6	36485	2.4E-11	0.472	yes	0.999	0.165	0.116	0.155	0.196	0.114	0.085	A	-0.064	0.194	0.357	0.233	0.141
rs10946808	A	G	HistoneClass1,ButyrophilinGenes	36485	5.8E-10	0.718	yes	0.999	0.484	0.128	0.000	0.511	0.125	0.000	G	-0.457	0.173	-1.009	0.306	0.812
rs2844479	T	C	HLAclassIII	33992	8.9E-09	0.670	yes	0.997	0.259	0.125	0.038	0.308	0.134	0.022	T	0.036	0.283	0.401	0.283	0.378
rs185819	T	C	HLAclassIII	36485	3.2E-08	0.519	yes	0.999	0.068	0.115	0.553	-0.107	0.125	0.391	T	0.038	0.206	0.135	0.231	0.860
rs1776897	C	T	HMGA1,LBH	36485	1.4E-08	0.089	yes	0.990	0.799	0.204	0.000	0.732	0.201	0.000	T	-0.884	0.948	-1.675	0.929	0.926
rs4713858	G	A	ANKS1,TCP11,ZNF76,DEF6,SCUBE3	39509	3.5E-08	0.860	yes	0.999	0.495	0.166	0.003	0.362	0.165	0.028	G	-1.278	0.587	-0.472	0.570	0.002
rs3748069	A	G	GPR126	39509	4.5E-14	0.722	yes	1.000	0.521	0.128	0.000	0.344	0.440	0.435	A	0.151	0.315	0.822	0.308	0.198
rs798544	G	T	GNA12	39509	6.5E-15	0.721	yes	0.999	0.380	0.129	0.003	0.364	0.127	0.004	T	-0.649	0.172	-0.339	0.314	0.018
rs10958476	C	T	PLAG1,MOS,CHCHD7,RDHE2,RPS20,LYN,TGS1,PENK	36485	6.6E-08	0.236	yes	0.993	0.274	0.138	0.047	0.191	0.138	0.167	T	-0.561	0.380	-0.748	0.370	0.417
rs7846385	C	T	PXMP3,ZFHX4	39509	4.7E-08	0.296	yes	1.000	0.164	0.128	0.201	0.182	0.126	0.147	T	0.751	0.310	0.218	0.307	0.001
rs4743034	A	G	ZNF462	39509	2.1E-08	0.238	yes	0.999	0.023	0.135	0.865	0.015	0.132	0.909	G	-0.258	0.360	-0.203	0.350	0.482
rs8756	C	A	HMGA2	39509	1.8E-16	0.497	yes	0.999	0.379	0.116	0.001	0.070	0.434	0.872	A	-0.360	0.202	-0.758	0.232	0.905
rs7153027	A	C	TRIP11,FBLN5,ATXN3,CPSF2	33992	1.1E-10	0.568	yes	1.000	0.021	0.116	0.857	0.338	0.190	0.075	A	0.202	0.222	0.084	0.237	0.340
rs4533267	A	G	ADAMTS17	39509	3.3E-08	0.283	yes	0.999	0.103	0.130	0.425	0.133	0.127	0.294	G	-0.307	0.321	-0.332	0.315	0.487
			CRLF3,ATAD5,CENT																	

Table 4.3: Continued

rs3760318	C	A	A2,RNF135	33992	1.8E-09	0.648	yes	0.997	0.405	0.121	0.001	0.363	0.119	0.002	A	-0.404	0.176	-0.810	0.265	0.994
rs4794665	A	G	NOG,DGKE,TRIM25, COIL,RISK	39509	9.9E-08	0.514	yes	0.998	0.303	0.117	0.010	0.281	0.115	0.014	A	1.047	0.205	0.649	0.234	**0.000**
rs757608	T	G	BCAS3,NACA2,TBX2, TBX4	39509	6.3E-08	0.343	yes	0.998	0.297	0.121	0.014	0.257	0.119	0.031	G	-0.129	0.266	-0.516	0.267	0.479
rs4800148	A	G	CABLES1,RBBP8,C18orf45	36485	3.7E-09	0.808	yes	0.999	0.430	0.145	0.003	0.488	0.142	0.001	A	-0.105	0.430	0.468	0.416	0.186
rs967417	C	A	BMP2	39509	1.5E-08	0.550	yes	0.999	0.207	0.116	0.075	0.238	0.114	0.037	C	0.296	0.216	0.429	0.234	0.623
rs724016	G	A	ZBTB38	33518	8.3E-22	0.485	no	0.997	0.491	0.116	0.000	28.826	30.246	0.341	A	-0.602	0.206	-0.988	0.233	0.512
rs1042725	C	T	HMGA2	29425	2.7E-20	0.515	no	0.997	0.377	0.117	0.001	0.305	0.437	0.485	T	-0.444	0.198	-0.751	0.233	0.679
rs4896582	G	A	GPR126	30111	2.4E-18	0.711	no	0.985	0.519	0.127	0.000	0.204	0.438	0.641	A	-0.677	0.175	-0.825	0.301	0.187
rs6060369	C	T	GDF5-UQCC	29425	1.4E-16	0.403	no	0.990	0.500	0.119	0.000	0.494	0.117	0.000	T	-0.869	0.237	-1.119	0.246	0.072
rs1492820	A	G	HHIP	33518	1.2E-11	0.549	no	0.995	0.339	0.118	0.004	0.343	0.127	0.007	A	0.604	0.218	0.728	0.238	0.149
rs8007661	C	C	TRIP11-ATXN3	23624	5.5E-10	0.599	yes	0.885	-0.139	0.123	0.259	-0.461	0.201	0.022	C	-0.184	0.256	-0.290	0.254	0.839
rs314277	A	C	LIN28B	29425	1.1E-08	0.147	yes	0.999	0.235	0.163	0.150	0.203	0.160	0.204	C	-0.083	0.567	-0.346	0.551	0.780
rs12986413	T	A	DOT1L	29425	2.9E-08	0.471	no	0.987	0.182	0.117	0.119	0.215	0.114	0.061	T	0.183	0.196	0.365	0.234	0.998
rs2562784	G	A	SH3GL3-ADAMTSL3	23624	6.4E-08	0.216	yes	0.999	0.255	0.140	0.068	0.057	0.153	0.708	A	-0.197	0.398	-0.469	0.386	0.876
rs9650315	G	T	CHCHD7-RDHE2	26003	3.8E-07	0.885	no	0.891	0.707	0.192	0.000	0.675	0.193	0.000	G	0.115	0.820	0.899	0.793	0.458
rs2040494	T	C	CDK6	29425	3.8E-07	0.535	no	0.992	0.439	0.117	0.000	0.269	0.154	0.080	C	-0.199	0.193	-0.914	0.235	0.121

* linear model $height \sim \beta_s\, sex + \beta_a\, age + \beta_B\, (P_{AB} + 2\,P_{BB})$, where P_{AB} and P_{BB} are the estimated probabilities of the AB and BB genotypes, respectively

** linear model $height \sim \beta_s\, sex + \beta_a\, age + \beta_{B1}\, (P_{AB1} + 2\,P_{BB1}) + \beta_{B2}\, (P_{AB2} + 2\,P_{BB2}) + \ldots + \beta_{B54}\, (P_{AB54} + 2\,P_{BB54})$, where P_{ABi} and P_{BBi} are the estimated probabilities of the AB and BB genotypes at i-th locis (i from 1 o 54)

*** linear model $height \sim \beta_s\, sex + \beta_a\, age + \beta_{AB}\, P_{AB} + \beta_{BB}\, P_{BB}$, where P_{AB} and P_{BB} are the estimated probabilities of the AB and BB genotypes, respectively

**** *p*-value from comparison between model (*) and (***) using Likelihood Ratio Test on 1 d.f.

Figure 4.17: **ROC plots showing the discriminative ability of the genetic weighted risk score, environmental factors and genetic weighted risk score and environmental factors combined for six extreme groups of TC in ERF.** The top three panels show ROC curves for the lower extreme groups, the bottom three panels show ROC curves for the upper extreme groups.

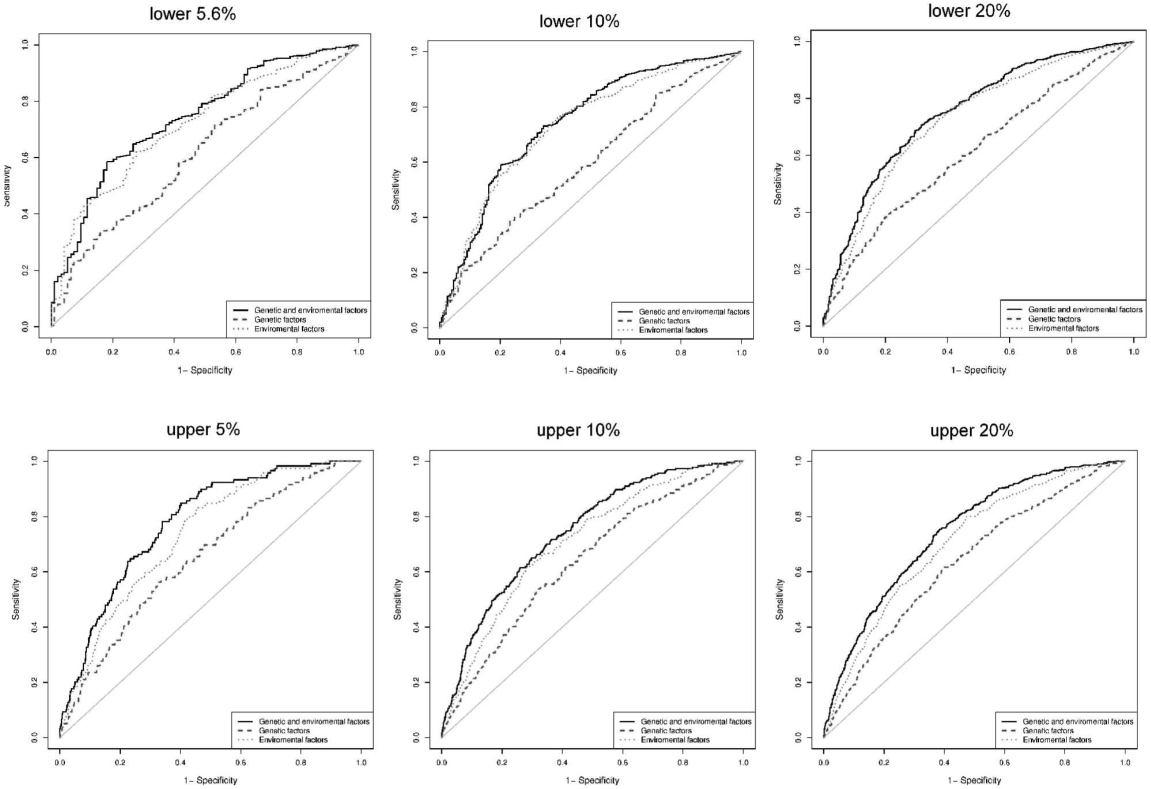


Figure 4.18: **ROC plots showing the discriminative ability of the genetic weighted risk score, environmental factors and genetic weighted risk score and environmental factors combined for six extreme groups of TC in RS.** The top three panels show ROC curves for the lower extreme groups, the bottom three panels show ROC curves for the upper extreme groups.

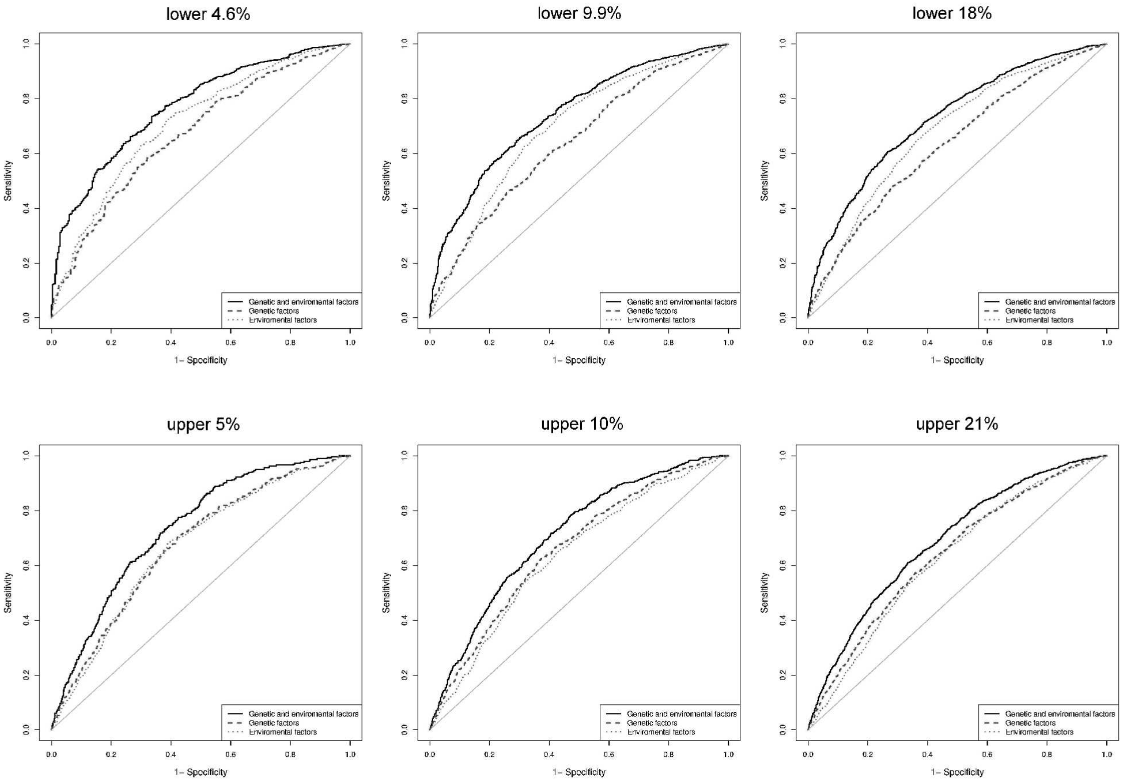


Table 4.4: **Minor allele frequencies and imputation quality values of the single nuclear polymorphisms used in the analysis**

	ERF			RS		
SNP name	MAF*	Quality**	R^2***	MAF*	Quality**	R^2***
rs2255141	0.27	1.00	1.00	0.28	1.00	1.00
rs10832963	0.22	0.99	0.97	0.25	0.98	0.95
rs174550	0.30	1.00	1.00	0.33	1.00	1.00
rs964184	0.13	0.99	0.97	0.13	1.00	0.98
rs7941030	0.42	1.00	0.99	0.38	1.00	1.00
rs11220463	0.14	0.95	0.86	0.11	0.96	0.85
rs11065987	0.42	0.96	0.94	0.41	1.00	1.00
rs1169288	0.37	0.97	0.95	0.30	0.98	0.97
rs1532085	0.38	0.97	0.95	0.39	1.00	1.00
rs3764261	0.39	0.97	0.94	0.33	1.00	1.00
rs2000999	0.22	0.99	0.97	0.19	1.00	0.99
rs7206971	0.49	0.99	0.99	0.46	0.99	0.99
rs7239867	0.14	1.00	1.00	0.17	1.00	1.00
rs6511720	0.17	0.97	0.92	0.12	1.00	1.00
rs10401969	0.06	0.97	0.82	0.07	0.98	0.85
rs4420638	0.24	0.82	0.57	0.20	0.83	0.53
rs492602	0.43	0.96	0.92	0.46	0.99	0.98
rs12027135	0.46	1.00	0.99	0.46	0.99	0.98
rs2479409	0.27	0.80	0.56	0.31	0.86	0.74
rs3850634	0.34	1.00	1.00	0.36	1.00	1.00
rs7515577	0.19	1.00	0.99	0.21	1.00	1.00
rs629301	0.22	0.98	0.94	0.23	1.00	1.00
rs2807834	0.29	0.90	0.82	0.34	0.93	0.88
rs514230	0.49	0.98	0.97	0.46	1.00	0.99
rs2277862	0.12	1.00	1.00	0.14	1.00	1.00
rs2902940	0.28	0.99	0.98	0.31	1.00	1.00
rs4297946	0.48	0.97	0.97	0.46	0.99	0.98
rs1800961	0.03	0.99	0.89	0.04	1.00	0.99
rs1367117	0.27	0.94	0.88	0.31	0.94	0.90
rs1260326	0.34	0.99	0.98	0.39	0.97	0.96
rs4299376	0.28	0.95	0.90	0.30	1.00	1.00
rs6759321	0.30	0.98	0.96	0.29	0.99	0.98
rs2290159	0.22	1.00	1.00	0.22	1.00	1.00
rs12916	0.38	0.98	0.97	0.37	0.97	0.96
rs6882076	0.33	0.99	0.98	0.36	0.96	0.95
rs3757354	0.19	0.99	0.97	0.21	1.00	1.00
rs1800562	0.07	1.00	0.98	0.06	1.00	1.00
rs3177928	0.14	1.00	1.00	0.15	1.00	1.00
rs2814982	0.08	0.98	0.91	0.10	0.98	0.93
rs9488822	0.34	0.97	0.95	0.30	1.00	0.99
rs1564348	0.16	0.99	0.96	0.15	1.00	1.00
rs2285942	0.18	0.95	0.87	0.15	0.98	0.93
rs2072183	0.29	0.76	0.53	0.23	0.89	0.74
rs2126259	0.07	1.00	0.98	0.10	1.00	1.00
rs1961456	0.33	0.95	0.91	0.32	0.98	0.96
rs1030431	0.37	0.98	0.97	0.34	1.00	1.00
rs2737229	0.27	0.99	0.98	0.31	1.00	1.00
rs2954022	0.49	0.99	0.99	0.46	1.00	1.00
rs11136341	0.39	0.90	0.81	0.35	0.89	0.80
rs581080	0.19	0.97	0.92	0.20	0.98	0.95
rs1883025	0.27	0.92	0.84	0.25	0.94	0.87
rs651007	0.16	0.97	0.91	0.21	0.98	0.96

* Minor allele frequency.

** The average posterior probability for the most likely genotype.

*** The squared correlation between imputed and true genotypes.

5

CHAPTER: GENERAL DISCUSSION

Genome-Wide Association Studies (GWAS) play a crucial role in dissecting the heritability of common traits and diseases that have complex genetic architecture. However, the advances in this field so far raise the question of the "missing heritability": the SNPs identified in GWAS generally explain only a small part of heritability - proportion of phenotypic variation attributed to genetic factors. The work described in this thesis addresses this issue and suggests methodological and computational solutions that in one way or the other relates to the GWAS approach. In section 'GWAS methodology and beyond' I begin with a description of the methodology and a software tool for GWAS, a large part of which I implemented in the `ProbABEL` package. `ProbABEL` is a software tool for GWAS which provides the mostly used instruments such as analysis of quantitative, binary, and time-to-event outcomes. Then, I address the missing heritability issue and describe two novel methods. One of them allows studies of frequent loss-of-function alleles in genome wide association studies which, besides the homozygous state, influences the phenotype in the compound heterozygote state, where I contributed to the data analysis and software development. Another method was proposed and mostly developed by me and allows detection of genetic variants which influence common traits in interaction with other genetic or environmental factors. I describe the R package `VariABEL` which I wrote and which implements these methods. In the section 'Post-GWAS studies of complex genetic architecture of common traits' I discuss results of the application of different approaches for studying human traits, i.e., height (where I contributed to the planing of the study and analysis of the data), total cholesterol (conducted by me) and arterial stenosis (where I contributed to the analysis plan and data analysis).

GWAS METHODOLOGY AND BEYOND

After the first successful GWAS in 2004, the method rapidly gained popularity among many research groups studying common phenotypes. This stimulated an increased demand for novel methodological solutions and software tools for GWAS. `ProbABEL` (described in Chapter 2.1) is a software package that we developed as a response to this demand. This package facilitates GWAS of imputed data in a fast and memory-efficient manner. The most popular GWAS methods including analysis of quantitative, binary, and time-to-event outcomes are implemented in `ProbABEL`. For quantitative traits I implemented a fast mixed-model -based score test enabling genome-wide SNP analysis in family-based studies, studies performed in genetically isolated populations as well outbred human and animal populations.

Analysis of imputed genetic datasets was implemented in the `ProbABEL` package using an approach based on regression on estimated probabilities. A limited number of software packages were available for this type of analysis. `SNPTEST`, for example, implements a score test based on missing data likelihood [36] allowing the study of both quantitative and binary outcomes. MACH2QTL and MACH2DAT [18, 37] implement regression models on estimated probabilities for quantitative and binary traits, respectively, in a manner similar to ProbABEL. ProbABEL extends the functionality available in these packages by allowing analysis using the Cox proportional hazards model. Furthermore, another advantage compared to SNPTEST is that when testing the interaction of a covariate with SNPs, it also provides the value of the global significance test.

I implemented the two-step mixed model-based score test in ProbABEL which is an extension of the family-based association score test suggested by Chen and Abecasis [38], and in its logic, it is similar to the GRAMMAR, GRAMMAR-GC and GRAMMAR-Gamma tests described by Aulchenko *et al.* [39, 40, 41]. In the test procedure, the model is split into two parts (see the equation 6 in section "Two-step score test approximation to the mixed model" of Chapter 2.1), the first of which contains the effects of nuisance parameters, including random genetic effects, and the second includes the parameters of interest (i.e., SNP effects and SNP-interacting covariates). In the second step of the procedure estimation is performed based on the estimates obtained from fitting the first part. Strictly speaking, a test defined in this manner is correct if the distributions of covariates in the first and the second parts of the model are independent conditional on the estimated phenotypic variance-covariance matrix. This assumption is most likely to be true when the covariates included in the basic model are environmental ones, and thus are not expected to exhibit a conditional correlation with SNPs. However, when endogenous risk factors, such as body mass index, are included as covariates in the basic model, some SNPs are expected to exhibit covariance with this covariate. In such situations, the covariate should be included in the second step analysis. This, however, may violate the assumptions of the score test if the covariate explains a large proportion of trait variance. In such situation we expect that the test will become conservative and may be less powerful compared to the classical maximum likelihood analysis.

GWAS of millions of imputed SNPs using the mixed model-based score test in ProbABEL takes a few days for samples of a few thousands of people (on one core of a Sun Fire X4540 server with an AMD Opteron CPU, 3.00 GHz). However, the relationship between CPU time and the number of subjects is not linear; as the number of subjects reaches $5,000$ or more, the mixed-model based analysis will take too much time (weeks to months) when using a single core. A straightforward approach to solve this problem would be to use parallel computations. Still, the non-linear dependency of computational time on the number of subjects may become a major analysis bottleneck with larger and larger studies becoming available. Although the recently published GRAMMAR-Gamma method [41] is extremely fast and its computational time is linear with sample size, this comes at the cost of losing accuracy: indeed the authors recommend to use a ProbABEL-type of approach when the genetic structure of a sample is very complex.

Other software packages which implement similar mixed-model functionality and are suitable for GWAS are MERLIN [42], QxPak [43] and EMMAX [44]. In particular, MERLIN implements the two-step score test [38], which is equivalent to the test implemented in ProbABEL in the absence of covariates. QxPak is a flexible tool for mixed modeling of quantitative traits, which implements classical full Maximum Likelihood and Restricted Maximum Likelihood estimation procedures. EMMAX [44] utilizes methodologically the same approach as ProbABEL: estimation of pairwise genetic relationship, calculation of the covariance matrix of phenotypes that models the effect of genetic relatedness on the phenotypes, followed by a score test of association [38].

One of the important advantages of ProbABEL at the time it was released was that this package was the only one which allowed analysis of imputed genetic data in a fast manner. To date, analyzing of imputed data sets is a routine work and all the packages aimed to genetic association analysis support this feature. Since its advent, ProbABEL is a part of an open source initiative - GenABEL project [45] - which provides the opportunity for other individual researchers and research groups to contribute to the development of the project. Due to that fact, ProbABEL has flexible mechanisms for the further improvement, that is necessary for any successful software product. One of the example of using this advantage is the development of DatABEL package allowing storage and access to large genetic data set and subsequent usage of this package in studies conducted by ProbABEL. Currently, this allows processing large genetic data sets which became available for analysis after imputation using the recent reference panels and provide a flexible basis for development of the new software packages aimed to analyze large sequence data.

ProbABEL has been widely adopted and has been used for the analysis of various data sets, including genome-wide analysis of such various traits as gout [46], waist circumference [47], smoking initiation [48], height [49], and others. In total, the manuscript describing the methodology and ProbABEL package had been cited 126 times by 24 December 2012 according to http://scholar.google.nl.

The rapid development of genotyping techniques enables the high-density genetic mapping studies where attention is focused on sequencing the whole human genome or address particular genomic regions. This provides the possibility to study rare variants in a broad context similarly to common variants in GWAS. Such studies require new methodology and software tools. In Chapter 2.2 I focused on studying rare lost-of-function (LOF) variants. I described a novel method - the collapsed double heterozygosity (CDH) test - which we developed for testing LOF variants when they influence a trait in a heterozygous state. We demonstrated both theoretically and empirically by simulations that using the collapsed genotypes in GWA analysis is more powerful than single SNP analysis in detecting the presence of multiple LOF variants at a particular gene locus. In a genome scan of the red hair color phenotype this CDH analysis resulted in considerably more significant association signals than single SNP analysis at

MC1R. Besides *MC1R*, no other region of collapsed heterozygosity association with red hair was identified. The additional genotyping of two causal SNPs in *MC1R* confirmed a recessive model underlying this gain in statistical power. The generalizability of CDH mainly depends on the effect sizes and frequencies of causal alleles. We expect CDH is generalizable to some of the known examples, such as HFE and hemochromatosis, where both the allele effect sizes and frequencies are similar to those in *MC1R*. Further, through simulations we showed our method is capable to find LOF alleles with smaller effect sizes ($GRR > 3$) but not with frequencies lower than 1%. It should therefore be emphasized this approach still requires causal alleles to be at some appreciable frequency ($> 1\%$) to be effectively tested and probably not useful for exceptionally rare variants unless they combine with common variants.

Another approach proposed for the identification of new genetic variants is testing for gene-gene and gene-environment interactions [19]. In Chapter 2.3, I described a novel approach for searching for interacting loci. I evaluated the type I error and the power of variance heterogeneity analysis for the detection of potentially interacting SNPs in the scenario when the interaction variable is unknown. Through simulations, I studied three different statistical tests of variance heterogeneity and showed that Levene's (Brown-Forsythe) test has an appropriate type I error and power. In a similar work of Paré at al [24] the same method was chosen for studying interactions. The main peculiarity of the variance heterogeneity method of interaction testing lies in the fact that there is no need to know the interacting factor(s) itself. The method allows detecting genetic variants involved in interaction(s) with this factor(s). However, I show that the power to detect an interacting SNP depends on the magnitude of the effect size of the factor with which this variant is interaction. Thus, for a wide range of designs, models and test, the absence of significant heterogeneity of variances cannot be interpreted as absence of interaction. However, a clear drawback of the method is that the specific interaction underlying the inflation remains to be determined. These specific interactions in particular on the gene-environment level are highly relevant for prediction and prevention of the disease.

Although our model assumes a SNP having additive effect and following Hardy-Weinberg distribution, and an interaction factor following a normal distribution, the same principal result—non-monotonic dependence of the power of the variance test on the main effect of interacting variable—should hold for other models as well as other types of interacting factor (e.g., binary or three-level, such as other SNPs). Moreover, a deviation from Hardy–Weinberg Equilibrium does not affect our major conclusions.

I implemented the method in the software package `VariABEL` which is described in Chapter 2.4. I also extended the variance heterogeneity method to imputed SNPs. The method I suggest, SVLM, is based on linear regression, which makes results obtained in individual studies easily meta-analyzable using conventional methods and tools. I have utilized the fact that the variance is, by definition, the expectation of squared values of the centered variable. This allowed us to re-formulate the task of estimation and analysis of variances of the trait as a task of regression analysis of the transformed trait. In this setting, methodological and computational tools developed for GWAS are applicable for the variance analysis.

In a frame of dissecting the heritability of complex traits, the method of interaction testing through genotypic variance heterogeneity measurement is gaining wide popularity. In Ref. [50], L. Rönnegård and W. Valdar summarize and compare the recently developed statistical methods for detecting loci involved in interaction and affecting phenotypic variability. The methods are classified in three groups: classical non-parametric methods, the full parametric methods modeling mean and variance and the two-stage approximations to parametric methods.

The methods described in Chapter 2.3, Ref. [24] and Ref. [51] are classical non-parametric. I (Chapter 2.3) and Paré *et al.* [24] propose to use Levene's test for variance heterogeneity testing. The work described in Ref. [51] suggests using individual P-value threshold for each SNP tested by Levene's test that may increase power under a variety of interaction scenarios. These methods require that the data can be grouped into genotype classes that makes them inapplicable on imputed data where genotypes are known with some probability. Another disadvantage is that these methods do not naturally provide inclusion of covariates such as sex, age and so on. Although the first difficulty can be overcome by using best-guess genotypes and the second one - by pre-adjusting the trait before the actual analysis, doing so potentially decrease the power of the analysis.

The second class of methods - the full parametric methods - consists of a Markov Chain Monte Carlo (MCMC) method [52] and the Double Generilized Linear Model (DGLM) [53]. The first method simultaneously estimates effects of the mean and variance on phenotype to fit the regression model. The second method is a deterministic classical estimation which consists of two steps. First, a model accounting for the mean difference is fitted, then, during the second step, a model is fitted that accounts for the variance difference by using the estimated squared residuals from the first step. Subsequently the model parameters from the second step are used to update the model parameters from the first step. After a number of cycles the best fitting parameters are estimated.

Since the full parametric methods are computationally demanding it motivated the development the two-stage approximations. P. Visscher and D. Posthuma considered in their work [54] the possibility to test interaction by variance heterogeneity measurement. They used the term 'environmental sensitivity' which in my work I relate to the interaction between genetic and environmental factors. In the last paragraph of the paper's discussion section they consider the possibility of using the squared values of the residuals adjusted for the fixed effects and covariates for testing interacting loci in GWAS. Independently, we proposed the approach (implemented in the `VariABEL` package and described in Chapter 2.4) called SVLM which is based on the idea which is similar as the one proposed by of P. Visscher and D. Posthuma [54]. Later, J. Yang applied the method of testing genotypic variance heterogeneity in a large GWAS of $\approx 170,000$ individuals which revealed *FTO* gene associated with BMI variability [25]. In this study, squared residuals adjusted for environmental covariates but not for the main SNP effect were used. Theoretically, this results in increased type I error and lower power, however, it was demonstrated empirically that this has only minor effect and does not affect the conclusions of the study.

L. Rönnegård and W. Valdar compared my SVLM method with the DGLM through simulations and showed that with a sample size of $10,000$ and normally distributed phenotype, both methods produce almost identical P-values (correlation 0.9996) and have similar power [50]. The DGLM is the method of choice if the accurate estimation and the extensive study of variance is required. Despite the fact that the SVLM is only an approximation of the DGLM-like approach, the main advantage is that my method is relatively fast: in fact, the SVLM demands only double the CPU time required for a regular GWAS under the same conditions (e.g., study sample size, CPU productivity) making it a fast alternative to the DGLM.

POST-GWAS STUDIES OF THE COMPLEX GENETIC ARCHITECTURE OF COMMON TRAITS

In Chapter 3.1 the predictive power of GWAS findings using human height as an example and compared it to Victorian approach was evaluated. It was demonstrated that the selected 54 loci previously found in GWAS all together explain 4%–6%, whereas Galton's mid-parental values explain 40% of the height variation. Adding genomic information to the mid-parental values provide only a small (1.3%) increase in the proportion of variance explained. In forensics and human medicine, the question of binary classification of a person (e.g., "very tall" or not) on the basis of some profile (score) is of high interest. To estimate the accuracy of the prediction of large values of height the ROC curve [55, 56] was used. In medicine, ROC analysis is used extensively for the evaluation of diagnostic tests. We show that the 54-loci genomic profile had a relatively low discriminative accuracy compared to what is necessary for an effective prediction ($AUC = 65\%$ for a person falling into the category of the 5% tallest).

In epidemiology, most of prediction studies are based on finding risk factors responsible for disease (such as age, gender, smoking, blood pressure and so on) and constructing a risk score which represents the joint influence of all known risk factors. Risk scores provide an estimation of the probability to develop the disease by a particular person. In our study (Chapter 3.1) we used genetic risk score which contained the associated genetic variants that reached the genome-wide significance threshold in GWAS served as risk factors. Recently, it was demonstrated that genetic variants showing GWAS signals below the genome-wide significance threshold explain additionally up to 2.7% of total cholesterol, 2.6% of low density lipids, 4.8% of high density lipids [57], 1% of depression [58] and 10% of height variation [30]. This indicates that there are many genetic risk variants with smaller effects which were undiscovered in GWAS because of the power and sample size issues and those variants can be successfully used for prediction.

The studies of variants with low GWAS signals were motivated by the work on height [31] where it was shown that 300,000 common SNPs all together could potentially explain 45% of height variance and that the remaining heritability is due to variants of lower frequency. This is concordant with the hypothesis that common traits are explained by many common SNPs each having a small effect size (the "infinitesimal model" [59]). It suggests that future GWASs with larger sample sizes will reveal novel associated genetic variants.

GWAS is a successful tool which will likely reveal many new associations between phenotypes and common genetic variants in the future studies. However, besides looking at common variants, it is also important to known the extent to which rare variants contribute to the heritability of common traits. In Chapter 3.2, the ability of known common genetic variants and environmental factors to discriminate extreme levels of total cholesterol (TC) was studied. One can see from Figure 3.4 which illustrates the relationship of total cholesterol categories vs. the mean number of risk alleles of common SNPs, that the number of risk alleles is slightly increased in the lower 5% TC group and slightly decreased in the upper TC group. This means the some of individuals from those groups are there due to factors other than known common variants. This can be due to the presence of individuals having lipid lowering therapy who were mistakenly included in the analysis. Alternatively, those factors can be rare environmental factors with large effect sizes or, more likely, rare genetic variants with large effects. The authors of a similar work on height [60] came to this conclusion about very short individuals. They studied the 1.5% of tallest and shortest individuals from a sample of 78,000 individuals and showed that some of the short stature is explained by factors other than common genetic variation. In Chapter 3.2 and in Ref. [60] the deviation from a purely polygenic model was observed only for very extreme phenotypic values. A possible explanation is that the factors due to which this deviation in height is observed (e.g. rare genetic variants with relatively large effect) explain a relatively small proportion of the heritability. The same picture has been observed in the work on hypertriglyceremia [34] in which the authors found that despite a significant genetic burden in patients with hypertriglyceremia those variants explain only 1.1% of the total trait variation (whereas clinical variables explained 19.7% and common genetic variants explained 20.8%).

Further study of extreme TC values is needed. It would be interesting to conduct a study where TC is simulated according to the model in which rare variants with relatively large effects determine a part of heritability and compare the picture of the discriminative ability of common genetic variants for this simulated trait with real TC values. Such study will demonstrate how sensitive the method is for detecting rare variants and possibly a proportion of such variants in TC.

It would be interesting to apply the approach used in the work on height [60] where a mean height values among extremes was compared to the simulated values to TC. This approach may actually give an approximate answer on the magnitude of the proportion explained by the factors other than common genetic variants.

Implications of findings from the study described in Chapter 3.2 are important for future studies of TC. There are multiple discussions on whether GWAS – which is designed to detect common variants with relatively small effect sizes – will serve as a productive instrument for this. My results suggest that the proportion of heritability attributed to common variants with small effects is simillar across the distribution. This suggests GWAS as an appropriate tool for future studies on TC, even for the search of those extreme values.

The question about frequencies and effect sizes of genetic variants responsible for variation in common traits and diseases has been under considerable discussion for a long time. The first genetic studies which located genetic regions associated with phenotypic variation, due to the specificity of the methods used (i.e., linkage analysis), discovered rare genetic alleles with large effect sizes on the traits under study. Some argue that the simplest genetic model which could be formed in the light of those discoveries at that time was a genetic model where many rare genetic variants with relatively large effects influence common traits (*common disease/multiple rare variants* hypothesis). However, later, GWAS showed that common traits for a large part are explained by common variants (*common disease / common variant* hypothesis).

The emergence of these two genetic models was preconditioned by a history of development of technical and methodological instruments of massive genotyping and analyzing the date and reflect only extreme manifestations of real genetic models underlying common traits. To date, many rare variants with large effect sizes and common variants with small effect sizes have been discovered for various common traits. Even based only on knowledge about the validity of these two models it is reasonable to assume that there should be variants with frequencies and effects sizes between those extreme cases. In fact, such variants have been discovered to date. Another model is a model where many rare variants with small effects influence traits. GWAS has limited power to detect such variants implying the need for a very large sample size (say tens of millions of individuals) will likely detect some of such variants. However, recent mutations in a population which may be present in one person only are difficult to detect in GWAS. Based on logical reasoning, the trait's associated variants could exist in the whole range of allelic frequency and effect sizes and the main question to answer in the future is what the amount of proportion is explained by each of the genetic model.

Studying rare variants in broad context was impossible until recently as the chips used for genotyping were designed for GWAS, and, therefore, addressed only common variants with allele frequency greater than 5%. The cost of sequencing is decreasing, however, it is still expensive to sequence large populations. A cheap alternative is a custom designed chip offered by Illumina and Affymetrix where only specific variants are genotyped. As an example, the chip which allows capturing exome sequencing provides genotyping of more than 1 million non-synonymous variants most of which are rare [61]. This chip may facilitate studies aimed to discovering causal variants as many of them are found in the codding regions. It covers 92.8% of the well-annotated genes, allowing almost complete whole exome sequencing. The Immunochip allows fine mapping of loci (about 50,000 variants) associated with vascular and inflammatory disease [62]. We used Illumina Metabochip [35] (which was designed for high-density genotyping of loci associated with metabolic traits) for studying craniocervical stenosis (Chapter 3.3).

We found SNPs in the *GLIS*3 gene associated with increased risk of craniocervical stenosis (OR 2.2-fold, 95% CI: 1.6–3.1). The SNP showing the strongest association signal is not present on common GWAS arrays illustrating the potential of the Metabochip as useful instrument for the detection associations. *GLIS*3 codes for the *GLIS* family zinc finger 3, which was found to play a critical role in the development of pancreatic β cells and insulin gene expression [63]. Deficiency of the *GLIS*3 product leads to hyperglycemia and hypoinsulinemia, as well as reduced β cell response to glucose [63, 64]. Previous genetic studies have found SNPs in the *GLIS*3 gene to be associated with increased fasting glucose levels [65] and diabetes mellitus type 1 and 2 [66, 67]. In the analysis we did not find an association between stenosis and loci previously associated with diabetes mellitus, even though diabetes mellitus has been shown to increase the risk of stenosis and high-grade stenosis in the carotid artery [68, 69]. The effect of the genetic variants discovered in this study is large compared to the effects generally discovered in GWASs motivating future similar studies.

This is the first study on craniocervical stenosis using a chip containing regions associated with metabolic traits and showing presence of association with a SNP which is not present on the chips commonly used in GWAS. The associated regions are responsible on elevated fasting glucose levels that is in correspondence with the previous studies on carotid intima-media thickness (IMT) which demonstrated that development of increased IMT after ten years of follow-up was associated with higher levels of fasting glucose at baseline [70, 71].

Validation of these findings in the replication cohort from a different ethnic population is the best approach, however, there was no replication cohort found by us. Cross-validation and bootstrapping can serve as a proxy for replication in a different cohort. However, this approach can not replace a replication: a magnitude of the GWAS signal, which in our study is slightly above genome-wide significance threshold, will perhaps be reduced after cross-validation and bootstrapping that make the findings insignificant even if there is a possible true association.

In this study, we observed high ethnic heterogeneity which may be a possible genetic confounder in the analysis and consequently could be the reason for the GWAS signal. The sample in this study consists of individuals from many different ethnic groups (and probably individuals with parents from different ethnic groups). Moreover, the high heterogeneity did not allow us to conduct separate analysis on the various genetically close subsamples. Consequently, I used principle component (PC) analysis with subsequent inclusion of three PCs in the regression model. Ten PC were chosen as these represented the genetic structure of population, whereas others PCs represented random noise. We used the inflation factor λ to control against possible inflation due to population stratification. In the study $\lambda \approx 1$, indicating negligibly small inflation.

future studies of large data sets / outlook

Genetic studies in modern genetic epidemiology involve processing the large data sets (such as millions of SNPs in thousands individuals) which requires appropriately fast software tools. For the last decade, there are tens of software tools developed for studying common and rare variants, GxG and GxE interaction, expression data, copy-number variations an so on. The increasing amount and new kind of genetic data require improving existing tools and developing the new ones. New approaches can optimize and decrease cost of this process of software development.

A general problem in bioinformatics is that many projects develop their own solutions for common features. The function of storage and retrieving data is an example. Many developers prefer to develop their own formats of storage and corresponding application programming interface to retrieve genetic data. Often, those solutions are inflexible (can not be easily extended to the new genetic data, e.g., genome sequence) and work slowly. Furthermore, different data formats create unnecessary barriers for analysis of the data by different software (e.g., a user of `GenABEL` or `ProbABEL` has to convert the data to use it in the `PLINK`). All those disadvantages seem to be not so severe when dealing with genetic data currently used in GWAS (a few millions SNPs, a few thousands of individuals). However, analysis of sequencing data probably will uncover those pitfalls. It will be costly and difficult to store a data set of a few terabytes in different formats to use it by different software tools. The solution can be the development of unique protocol and software tools allowing optimal storage and fast access to the data. A good example of such software in elementary particles physics is `ROOT` [72] - a software package developed at CERN in 1995 which offers a flexible tool (the TTtree class) for data storage which is optimized to reduce disk space and enhance access speed. `ROOT` is the primary tool for processing data from the Large Hadron Collider's experiments estimated at several petabytes per year. Today, most experimental plots and results in the field are obtained using ROOT. For future software projects in genetics it would be useful to develop a similar tool (or adapt the existing ones) which is optimized for genetic data. Such projects are possible only in the framework of open source and collaborative initiatives. The GenABEL project [45] can server as a good example of such an initiative in statistical genetics. It provides a platform for development and maintenance of software and the exchange of ideas on an open source basis. The GenABEL project has already addressed the data storage issue in DatABEL [73], however, the future advances in the field will require further development of this tool.

GWAS substantially improved our understanding of the genetic architecture of many traits. Five years ago very little was known about genetic variants influencing variation of a heritable traits such as human height in the general population. Today, GWAS has discovered several hundred genetic variants which explain more than 10% of height variance. This is still too little for successful prediction, however it gives us valuable clues about the genetic architecture of common traits and raises the question about possible strategies for the future studies. It was shown (Ref. [74]) that the number of loci identified in GWAS is nearly proportional to sample size. This provides a good argument in favor of performing GWASs with increased sample sizes as it will allow to detect variants with smaller effects. By looking at the distribution of the effect sizes of the height variants discovered up to date, one can see a clear exponential shape - every next variant discovered in a new GWAS has a smaller effect size than the previous one. By fitting exponent to this distribution it was roughly estimated from fit parameters that a total number of height genetic common variants explaining 80% of height variation is about $100,000$ [75]. To discover this number of genetic variants, we need to assess 50 millions individuals in GWAS (assuming that the number of variants discovered is proportional to sample size). This is about three times the size of the population of The Netherlands or about two times the size of the population of Australia. Taking into account that the maximum sample size used in the modern GWASs does not exceed a few hundreds of thousands of individuals, revealing most of the genetic variants responsible for human height variation seems a distant prospect.

These calculations are made under assumption that the phenotype is explained by common variants (the infinitesimal model), however, there are arguments supporting the genetic model in which rare variants with relatively large effects influence the phenotype. One of them which came from the theory of evolutionary states that as a disease is deleterious to fitness, variants that promote it should be selected against and the existence of such variants reflects the balance between mutation rate and purifying natural selection [59]. In the case of quantitative phenotypes, this argument can be fully applied to the individuals with extreme phenotypic values as they are suffering more from selection pressure (e.g., those with high TC levels are more predisposed to cardiovascular diseases). There is evidence that only individuals with extreme phenotypes are possibly enriched with rare variants with relatively large effect [60]. Those individuals are potential candidates for the future sequencing studies which will lead to discovery of new rare variants.

A very important field in epidemiological studies is trait prediction. The common approach nowadays is using GWAS findings as predictors which through constructions of a risk score profile give us the predicted estimation of phenotypic values. The power of this approach when applied for common traits such as height is rather disappointing (only 10% of height variation can be explained by GWAS findings). There is a greater success in animal studies where the inbreeding values (traits of interest) of animals is predicted based on the relative genetic relatedness of individuals. This approach allows the prediction of traits with a high accuracy in inbred populations (which is common in animal studies), however, in outbred population (such as often used in GWAS) the accuracy drops drastically. Nevertheless, this methodology can be applied to future studies of human isolated populations (such as ERF) which have relatively high inbreeding.

GWAS is, undoubtedly, the tool of choice for future studies and will facilitate the discovery of many new common genetic variants. However, rare variants and variants interacting with other factors (both genetic and non-genetic) explain a fraction of the heritability of common traits. The core of this thesis consists of studies of the complex genetic architecture of common traits. This includes the development of GWAS methodology together with their implementation in software tools as well as applying these new approaches to the study of common and rare genetic variants and genetic variants involved in interactions. Application of these methods and their further improvement will result in new findings in future studies.

BIBLIOGRAPHY

1. Botstein, D, White, R. L., Skolnick, M & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32,** 314–331 (May 1980).
2. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics* **33,** 228–237 (Mar. 2003).
3. Roberts, R. J. Restriction endonucleases. *CRC Critical Reviews in Biochemistry* **4,** 123–164 (Nov. 1976).
4. Mullis, K. B. Target amplification for DNA analysis by the polymerase chain reaction. *Annales De Biologie Clinique* **48,** 579–582 (1990).
5. Hearne, C. M., Ghosh, S & Todd, J. A. Microsatellites for linkage analysis of genetic traits. *Trends in Genetics: TIG* **8,** 288–294 (Aug. 1992).
6. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (Sept. 1996).
7. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 2001).
8. Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291,** 1304–1351 (Feb. 2001).
9. Consortium, T. I. H. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (Oct. 2005).
10. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (Oct. 2007).
11. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (Sept. 2010).
12. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (Nov. 2012).
13. Matsuzaki, H. *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nature Methods* **1,** 109–111 (Nov. 2004).
14. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)* **308,** 385–389 (Apr. 2005).
15. *HuGENavigator|GWAS Integrator|Search* <`http://hugenavigator.net/HuGENavigator/gWAHitStartPage.do`> (visited on 31/01/2013).

16. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5,** e1000529 (June 2009).

17. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* **84,** 210–223 (Feb. 2009).

18. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34,** 816–834 (2010).

19. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

20. McQuillan, R. *et al.* Evidence of inbreeding depression on human height. *PLoS genetics* **8,** e1002655 (2012).

21. Vazquez, C *et al.* Thirteen cystic fibrosis patients, 12 compound heterozygous and one homozygous for the missense mutation G85E: a pancreatic sufficiency/insufficiency mutation with variable clinical presentation. *Journal of medical genetics* **33,** 820–822 (Oct. 1996).

22. Singleton, A. B., Hardy, J., Traynor, B. J. & Houlden, H. Towards a complete resolution of the genetic architecture of disease. *Trends in genetics: TIG* **26,** 438–442 (Oct. 2010).

23. Mukherjee, B., Ahn, J., Gruber, S. B. & Chatterjee, N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American Journal of Epidemiology* **175,** 177–190 (Feb. 2012).

24. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet* **6,** e1000981 (2010).

25. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490,** 267–272 (Oct. 2012).

26. Visscher, P. M. *et al.* Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American journal of human genetics* **81,** 1104–1110 (Nov. 2007).

27. MN, W., H, L. & CM, L. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40,** 575–583 (2008).

28. G, L., AU, J. & C, G. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40,** 584–591 (2008).

29. DF, G., GB, W. & G, T. Many sequence variants affecting diversity of adult human height. *Nat Genet* **40,** 609–615 (2008).

30. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (Oct. 2010).

31. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42,** 565–569 (July 2010).

32. F, G. Regression towards mediocrity in hereditary stature. *Journal of the anthropological institute* **15,** 246–263 (1886).

33. Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS genetics* **7,** e1002198 (July 2011).

34. Johansen, C. T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature genetics* **42,** 684–687 (Aug. 2010).

35. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics* **8,** e1002793 (Aug. 2012).

36. 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).

37. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10,** 387–406 (2009).

38. Chen, W.-M. & Abecasis, G. R. Family-based association tests for genomewide association scans. *Am J Hum Genet* **81,** 913–926 (2007).

39. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).

40. Amin, N., van Duijn, C. M. & Aulchenko, Y. S. A genomic background based method for association analysis in related individuals. *PLoS One* **2,** e1274 (2007).

41. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nature genetics* **44,** 1166–1170 (Oct. 2012).

42. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30,** 97–101 (2002).

43. Perez-Enciso, M. & Misztal, I. Qxpak: a versatile mixed model application for genetical genomics and QTL analyses. *Bioinformatics* **20,** 2792–2798 (2004).

44. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42,** 348–354 (Apr. 2010).

45. *GenABEL.org GenABLE'ing genetical research* http://www.genabel.org/. <`http://www.genabel.org/`> (visited on 18/12/2012).

46. Woodward, O. M. *et al.* Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *Proc Natl Acad Sci U S A* **106,** 10338–10342 (2009).

47. 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).

48. Vink, J. M. *et al.* Genome-wide association study of smoking initiation and current smoking. *Am J Hum Genet* **84,** 367–379 (2009).

49. Estrada, K. *et al.* A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. *Hum Mol Genet* **18,** 3516–3524 (2009).

50. Rönnegård, L. & Valdar, W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC genetics* **13,** 63 (2012).

51. Deng, W. Q. & Paré, G. A fast algorithm to optimize SNP prioritization for gene-gene and gene-environment interactions. *Genetic epidemiology* **35,** 729–738 (Nov. 2011).

52. Sorensen, D. Developments in statistical analysis in quantitative genetics. *Genetica* **136,** 319–332 (June 2009).

53. Rönnegård, L. & Valdar, W. Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics* **188,** 435–447 (June 2011).

54. Visscher, P. M. & Posthuma, D. Statistical power to detect genetic Loci affecting environmental sensitivity. *Behavior genetics* **40,** 728–733 (Sept. 2010).

55. AC, J., MC, P., EW, S. & van Duijn CM. Revisiting the clinical validity of multiplex genetic testing in complex diseases. *Am J Hum Genet* **74,** 585–588 (2004).

56. AC, J. *et al.* Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* **8,** 395–400 (2006).

57. Demirkan, A. *et al.* Genetic architecture of circulating lipid levels. *European journal of human genetics: EJHG* **19,** 813–819 (July 2011).

58. Demirkan, A *et al.* Genetic risk profiles for depression and anxiety in adult and elderly cohorts. *Molecular Psychiatry* **16,** 773–783 (2011).

59. Gibson, G. Rare and common variants: twenty arguments. *Nature reviews. Genetics* **13,** 135–145 (Feb. 2011).

60. Chan, Y. *et al.* Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. *PLoS genetics* **7,** e1002439 (Dec. 2011).

61. Jiang, T., Yang, L., Jiang, H., Tian, G. & Zhang, X. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. *Science China. Life sciences* **54,** 945–952 (Oct. 2011).

62. Keating, B. J. *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PloS one* **3,** e3583 (2008).

63. Kang, H. S. *et al.* Transcription factor Glis3, a novel critical player in the regulation of pancreatic beta-cell development and insulin gene expression. *Molecular and cellular biology* **29,** 6366–6379 (Dec. 2009).

64. Boesgaard, T. W. *et al.* Variants at `DGKB/TMEM195`, `ADRA2A`, `GLIS3` and `C2CD4B` loci are associated with reduced glucose-stimulated beta cell function in middle-aged Danish people. *Diabetologia* **53,** 1647–1655 (Aug. 2010).

65. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics* **42,** 105–116 (Feb. 2010).

66. Hu, C. *et al.* Variants from `GIPR`, `TCF7L2`, `DGKB`, `MADD`, `CRY2`, `GLIS3`, `PROX1`, `SLC30A8` and `IGF1` are associated with glucose metabolism in the Chinese. *PloS one* **5,** e15542 (2010).

67. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics* **41,** 703–707 (June 2009).

68. Göksan, B, Erkol, G, Bozluolcay, M & Ince, B. Diabetes as a determinant of high-grade carotid artery stenosis: evaluation of 1,058 cases by Doppler sonography. *Journal of stroke and cerebrovascular diseases: the official journal of National Stroke Association* **10,** 252–256 (Dec. 2001).

69. Inchiostro, S *et al.* Prevalence of diabetes and/or ischaemic heart disease in classes of increasing carotid artery atherosclerosis: an ultrasonographic study. *Diabetic medicine: a journal of the British Diabetic Association* **20,** 670–676 (Aug. 2003).

70. Folsom, A. R. *et al.* Relation of carotid artery wall thickness to diabetes mellitus, fasting glucose and insulin, body size, and physical activity. Atherosclerosis Risk in Communities (ARIC) Study Investigators. *Stroke; a journal of cerebral circulation* **25,** 66–73 (Jan. 1994).

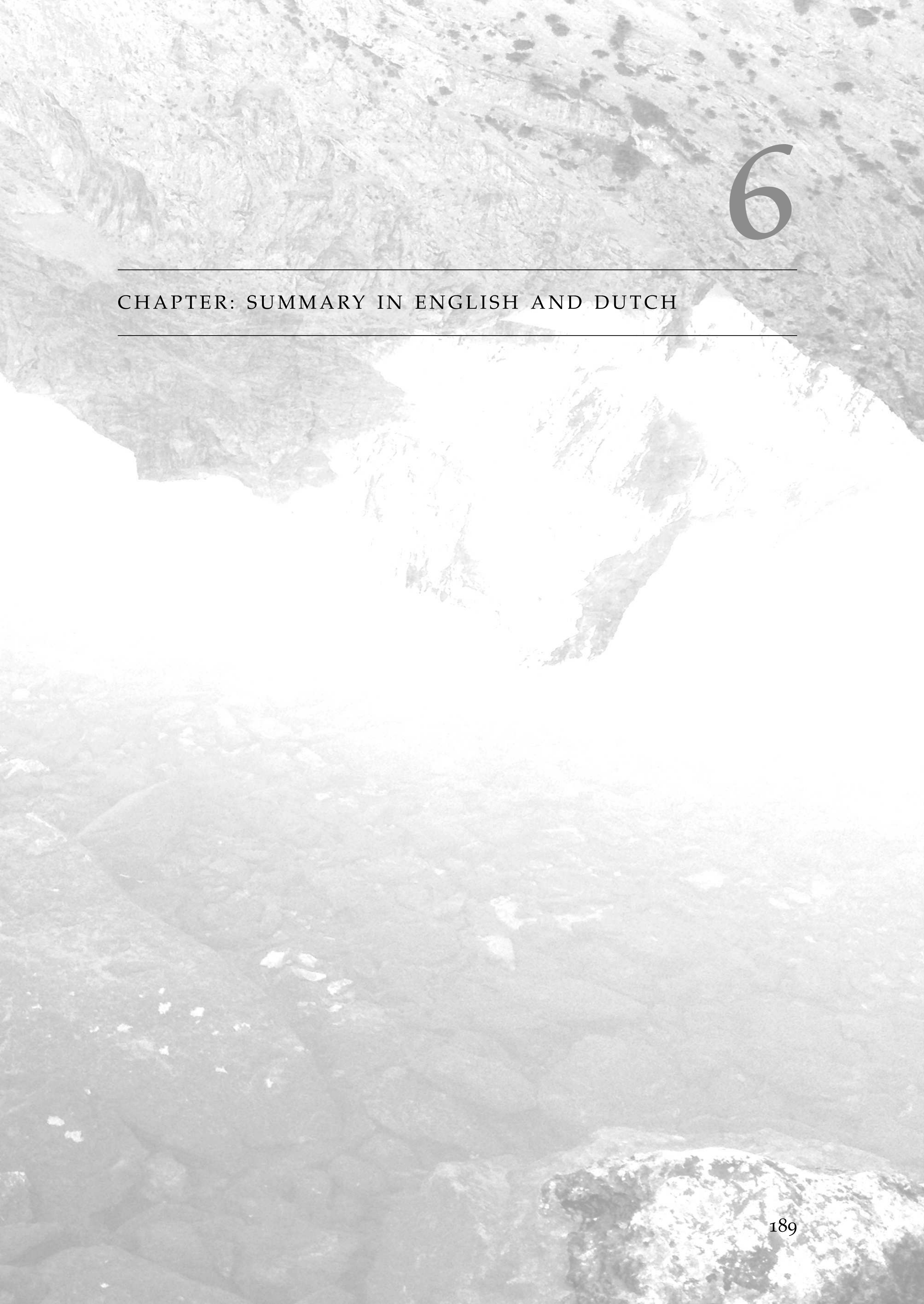
71. Tropeano, A., Boutouyrie, P., Katsahian, S., Laloux, B. & Laurent, S. Glucose level is a major determinant of carotid intima-media thickness in patients with hypertension and hyperglycemia. *Journal of hypertension* **22,** 2153–2160 (Nov. 2004).

72. *ROOT A Data Analysis Framework* http://root.cern.ch/drupal/. <`http://root.cern.ch/drupal/`> (visited on 17/12/2012).

73. *DatABEL package GenABEL.org* http://www.genabel.org/packages/DatABEL. <`http://www.genabel.org/packages/DatABEL`> (visited on 17/12/2012).

74. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90,** 7–24 (Jan. 2012).

75. Goldstein, D. B. Common genetic variation and human traits. *The New England journal of medicine* **360,** 1696–1698 (Apr. 2009).



6

CHAPTER: SUMMARY IN ENGLISH AND DUTCH

SUMMARY

Genome-wide association studies (GWAS) have substantially improved our understanding of the complex genetic architecture of many common traits. For the last decade, more than a thousand genetic variants were discovered using GWAS. The fast development of the field necessitated the improvement of existing instruments as well as the development of new ones. This thesis discusses methodology, software tools and new approaches facilitating the study of the complex genetic architecture of common traits.

In Chapter 2.1, we describe the software tool ProbABEL, which allows efficient running of GWASs of millions of SNPs in populations of thousands of individuals. We implemented the most popular GWAS methods including analysis of quantitative, binary, and time-to-event outcomes. For quantitative traits we implemented a fast mixed-model-based score test for association in samples with differential relationships that will facilitate analysis of family-based and genetically isolated populations. A very important feature of ProbABEL is the ability to analyze imputed SNPs. Analysis of millions of SNPs using the mixed-model-based score test in ProbABEL takes a few days for samples of several thousands of people (on a Sun Fire X4540 server with an AMD Opteron 2356 CPU, 2.30 GHz, using only one CPU core). We also implemented support for testing for interaction between genetic variants and environmental factors through regression analysis. Since its first release in 2009 ProbABEL has been widely adopted and has been used in the analysis of many data sets, for example in GWAS of traits like gout, waist circumference, smoking initiation, and height.

Genome-wide association studies were basically designed to test for associations with common genetic variants, however, there is evidence that rare variants with relatively large effect sizes can capture a sizeable fraction of the heritability. Particularly, various loss-of-function (LOF) variants have been shown to be responsible for trait variation. Multiple LOF variants at the same locus can act not only in the homozygous state, but also in the compound heterozygous (CH) state, where the presence of two different LOF variant alleles at the same gene, one on each homologuous chromosome, influences the phenotype. In Chapter 2.2, we demonstrate both theoretically and empirically by simulations that using such a genetic model can be more powerful than a regular GWAS approach where a single variant is tested. In a genome-wide scan of the red hair color phenotype this analysis resulted in considerably more significant association signals than single SNP analysis at *MC1R*. Besides *MC1R*, no other region of compound heterozygous association with red hair was identified. We expect that this test is generalizable to some of the known examples, such as *HFE* and hemochromatosis, where both the allele effect sizes and frequencies are comparable to those of the *MC1R* alleles.

Additionally, we addressed a genetic model in which a part of the heritability of common traits is explained by gene-gene and gene-environment interactions.

In Chapter 2.3, we describe a novel method which tests genetic variants for the presence of interaction. We propose to use Levene's variance homogeneity test for detecting genetic variants affecting phenotypic variation. We showed that such variants can influence a phenotype by interaction with other genetic or non-genetic factors. The main advantage of this method is that the interaction analysis does not require knowing the interacting factor (whether genetic or non-genetic). The method answers the question whether a given variant is involved in interaction or not.

In Chapter 3.1, we estimated the predictive power of 54 loci responsible for height variation and compared it to the method proposed by Galton more than a hundred years ago. We concluded that, compared to Galton's method, the predictive power of these genetic variants is low. Studying extremely tall individuals shows a similar picture: the 5% tallest individuals are discriminated from the rest of population with an AUC value of only 65%, which is low. This conclusion leads us to the issue of the hidden heritability and raises a question about further possible strategies for studying the genetic architecture of complex traits which will allow detecting new associations with genetic variants.

As mentioned earlier, there is evidence that rare variants with relatively large effects can be responsible for part of heritability. It is difficult to detect such variants either in linkage analysis (because the effect is not large enough) or in GWAS because of the low allele frequency. In Chapter 3.2 we studied the extent to which common variants with relatively small effects influence extreme levels of total cholesterol (TC) in two populations (i.e. the family-based Erasmus Rucphen Family study and the population-based Rotterdam Study). We studied the effect of common and rare variants implicated in risk factors for cardiovascular disease. The picture of the discriminative ability suggests that common variants have high predictive ability even in the extreme levels of total cholesterol.

There are examples demonstrating the existence of traits where both common and rare variants are located in the same loci as the GWAS signals. In Chapter 3.3, an association between carotid artery stenosis and loci harboring SNPs previously discovered in GWASs of metabolic traits is described. We discovered a variant in the *GLIS3* gene associated with carotid artery stenosis with 2.2 (95%CI $1.6 - 3.1$)-fold risk increase. This is a relatively large effect size compared to earlier GWAS findings and hence motivates future similar studies. However, the limitations of our study are its small sample size and lack of a replication cohort.

The core of this thesis are the methodology and accompanying software tools for studying both common and rare genetic variants as well as variants involved in interactions. The methods developed in the framework of this thesis show a high potential for the discovery of new genetic factors in future studies and some of them have already been instrumental in the discovery of new genetic variants.

SAMENVATTING

Genoomwijde associatiestudies (GWAS) hebben onze kennis van de complexe genetische structuur van een groot aantal veelvoorkomende kenmerken aanzienlijk verbeterd. In de afgelopen tien jaar zijn meer dan duizend genetische varianten gevonden met behulp van GWAS. De snelle ontwikkeling van het veld leidde zowel tot het verbeteren van bestaande hulpmiddelen als tot het ontwikkelen van nieuwe. In dit proefschrift worden methodologie, computerprogramma's en nieuwe benaderingen besproken die onderzoek naar de complexe genetische architectuur van veelvoorkomende kenmerken vergemakkelijken.

In hoofdstuk 2.1 beschrijven we het computerprogramma `ProbABEL` dat het mogelijk maakt om GWASen van miljoenen SNPs in populaties bestaande uit duizenden individuen op een efficiënte manier uit te voeren. We hebben de populairste GWAS methoden geïmplementeerd, waaronder analyse van kwantitatieve, binaire, en tijd-tot-gebeurtenis uitkomsten. Voor kwantitatieve eigenschappen hebben we een snelle, op mixed-models gebaseerde score-test voor associatie in samples met differentiële relaties geïmplementeerd, die analyse mogelijk maakt van op families gebaseerde populaties en genetisch geïsoleerde populaties. Een van de belangrijkste kenmerken van `ProbABEL` is de mogelijkheid om geïmputeerde SNPs te kunnen analyseren. Gebruik makend van de op mixed-models gebaseerde score test duurt de analyse van miljoenen SNPs voor een onderzoekspopulatie van enkele duizenden individuen enkele dagen (op een Sun Fire X4540 server met een AMD Opteron 2356 CPU, 2.30GHz, gebruikmakend van één processorkern). Onze implementatie ondersteunt ook het testen van interacties tussen genetische varianten en omgevingsfactoren door middel van regressieanalyse. Sinds de eerste editie in 2009 is `ProbABEL` op grote schaal in gebruik genomen en gebruikt voor de analyse van vele datasets, bijvoorbeeld voor de GWAS van kenmerken als jicht, heupomtrek, beginnen met roken en lengte.

Genoomwijde associatiestudies zijn in principe ontworpen om te testen voor associatie met veelvoorkomende genetische varianten. Er is echter bewijs voor het feit dat zeldzame varianten met een relatief groot effect een noemenswaardig deel van de erfelijkheid kunnen verklaren. Zo is aangetoond dat verschillende functieverliesvarianten (*loss-of-function*, LOF varianten) verantwoordelijk zijn voor variatie in kenmerken. Meerdere LOF varianten op dezelfde locus kunnen niet alleen samenwerken in de homozygote toestand, maar ook in de samengestelde heterozygote (*compound heterozygote*, CH) toestand, waarbij de aanwezigheid van twee verschillende LOF variant allelen in hetzelfde gen, een op elk homoloog chromosoom, het fenotype beïnvloeden. In hoofdstuk 2.2 tonen we zowel theoretisch als empirisch door middel van simulaties aan dat het gebruik van zo'n genetisch model krachtiger is dan de reguliere GWAS benadering waarbij een enkele variant wordt getest. In een genoomwijde scan van het fenotype "rode haarkleur" resulteerde deze analyse in aanzienlijk meer significante associatiesignalen dan de analyse van een enkele SNP in *MC1R*. Behalve *MC1R* werd geen ander gebied met samengesteld heterozygote associatie met rood haar geïdentificeerd. We verwachten dat deze test te generaliseren is voor enkele van de bekende voorbeelden als *HFE* en hemochromatose, waar zowel de effectgroottes als de frequenties van de allelen vergelijkbaar zijn met die van de *MC1R* allelen.

Daarnaast hebben we een genetisch model gemaakt waarin een deel van de overerfbaarheid van veelvoorkomende kenmerken verklaard wordt door gen-gen en gen-omgeving interacties. In hoofdstuk 2.3 beschrijven we een nieuwe methode welke genetische varianten test op de aanwezigheid van interactie. Wij stellen voor om Levene's variantiehomogeniteitstest te gebruiken voor het detecteren van genetische varianten die de variatie in fenotype beïnvloeden. We tonen aan dat zulke varianten een fenotype kunnen beïnvloeden door interactie met andere genetische en niet-genetische factoren. Het grote voordeel van deze methode is dat voor de interactieanalyse geen kennis nodig is over de interagerende factor (genetisch of niet-genetisch). De methode beantwoordt de vraag of een gegeven variant betrokken is bij interactie, of niet.

In hoofdstuk 3.1 schatten we het voorspellend vermogen af van 54 loci verantwoordelijk voor lengte en vergeleken dat met de methode welke meer dan honderd jaar geleden door Galton werd voorgesteld. We concludeerden dat, vergeleken met Galtons methode, het voorspellend vermogen van deze genetische varianten laag is. Onderzoek naar extreem lange individuen toont een vergelijkbaar beeld: de 5% langste individuen worden van de rest van de populatie onderscheiden met een AUC-waarde van slechts 65%, hetgeen laag is. Deze conclusie brengt ons bij het probleem van de onverklaarde erfelijkheid en roept vragen op over mogelijkheden voor strategieën voor verder onderzoek naar de genetische architectuur van complexe eigenschappen welke zullen leiden tot het detecteren van nieuwe associaties met genetische varianten.

Zoals eerder vermeld is er bewijs dat zeldzame varianten met relatief grote effecten verantwoordelijk kunnen zijn voor een deel van de erfelijkheid. Het is moeilijk om zulke varianten te detecteren met behulp van linkage analyse (omdat het effect niet groot genoeg is), of door GWAS omdat de allelfrequentie te laag is. In hoofdstuk 3.2 onderzochten we de mate waarin veelvoorkomende varianten met kleine effecten invloed uitoefenen op extreme waarden van totaalcholesterol (TC) in twee populaties (te weten de familiegebaseerde Erasmus Rucphen Familiestudie en de populatiegebaseerde Rotterdam Studie). We onderzochten in hoeverre bekende genetische varianten en omgevingsfactoren gebruikt kunnen worden om mensen met extreme TC-waarden van elkaar te kunnen onderscheiden. Onze analyse laat zien dat frequente genetische variatie een onderscheidend vermogen heeft over de gehele distributie.

Er zijn voorbeelden van het bestaan van kenmerken waarbij zowel veelvoorkomende als zeldzame varianten in dezelfde loci bevinden als de GWAS signalen. In hoofdstuk 3.3 wordt een studie beschreven waarin de associatie tussen stenose van de halsslagader en loci met daarin SNPs die eerder geassocieerd werden met metabole kenmerken. Wij vonden een variant in het *GLIS3* gen, welke geassocieerd werd met een 2.2 (95%CI $1.6 - 3.1$)-voudige toename in risico voor stenose van de halsslagader. Dit is een relatief groot effect vergeleken met eerdere GWAS resultaten en motiveert dus vergelijkbare studies in de toekomst. Ons onderzoek wordt echter gelimiteerd door een kleine studiepopulatiegrootte en het ontbreken van een replicatiecohort.

De kern van dit proefschrift wordt gevormd door de methodologie en de bijbehorende computerprogramma's voor het bestuderen van zowel veelvoorkomende als zeldzame genetische varianten, alsmede varianten die betrokken zijn bij interacties. De methoden die ontwikkeld zijn in het kader van dit proefschrift tonen een groot potentieel voor de ontdekking van nieuwe genetische factoren in toekomstige onderzoeken en sommige zijn al nuttig geweest voor de ontdekking van nieuwe genetische varianten.

I would like to acknowledge Cornelia van Duijn and Ben Oostra for being my 'promotors' and Lennart Karssen for supervising the final year of my PhD study. Dear Cornelia, thank you for giving me such a great opportunity to work in your group. It was extremely useful for me to work under your supervision. Dear Ben, thank you for your advices on my work. Dear Lennart, you supervised my work for almost a year, however, this was the crucial period for this thesis. Thank you for your time and everything you did for me. My thesis committee: Dear Franco Duran, Manfred Kayser and Andrea Uitterlinden, thank you for your valuable comments on my thesis. Dear Paul Eilers and Albert Hofman, thank you for being in my committee. My supervisor: Dear Yurii Aulchenko, it was a great pleasure to work under your supervision for the first almost four years. Your amazing skills in genetics, statistics, mathematics, programming and many others made it possible to do those interesting projects.

I would like to thank my colleagues. Dear Aaron, thank you for helping me with English language from the beginning of my work in the department: your comments on the grant proposal to the NVIDIA company was very important. Dear Carla, Elisa, Linda, and Sara thank you for sharing your knowledge and experience with me. Dear Najaf, thanks for your advices on my work and the discussions of the results. Dear Elza, you were always helpful and friendly. Dear Maarten, it was a pleasure to share a room with you especially due to your amazing sense of humor and openness. Dear Claudia, thank you for the nice parties. Dear Adriana, many thanks for the nice chats and beautiful Latin American music you showed to me. Dear Jeneatte, Benardetta, Andy, Andrea and Revanius it was a pleasure to work with you. Dear Natalia, I was always happy to see you around in ErasmusMC during the late evenings and sometimes even after midnight. I would like to thank Helena Shmidt and Reinhold Schmidt for giving me opportunity to analyze the ASPS data and work together. I special would like to thank my colleagues from Laboratory of Genetic Recombination and Segregation, Novosibirsk Institute of Cytology and Genetics. Dear Tatiana Axenovich, Pavel Borodin, Nadezhda Belonogova, Irina Zorkoltseva, Kirichenko Anatoly, Gulnara Svishcheva, Anna Torgasheva and others thanks for being such nice people, it was a pleasure to work with you. Dear Hilde Peeters, I was always glad to see you in the department during your short visits. Dear Evita van den Herik, thank you such a good teamwork on the stenosis project. Dear Fan, it was a pleasure to work with you. Dear Abbas Dehghan, thanks for all your help. Dear Ayse, I am grateful to you for everything. Thank you for being a good friend and helpful colleague. I wish you all the best.

I would like to acknowledge my supervisors from laboratory 3-2 of Budker's Institute of Nuclear Physics where I did my Bachelor's and Master's studies. Dear Ivan Nikolaev, I am very grateful for all the time you spent on teaching me. Besides physics, you have deep knowledge in many other different fields and it was always interesting to talk with you. There are actually a lot of things that I learned from you and which were important for my work described here. Dear Vladimir Evgenievich Blinov, thank you very much for giving me such a great opportunity to work in your laboratory. I feel lucky that I had chance to work under your supervision. Your professionalism, talent in organizing work and amazing teaching skills was crucial for my scientific and personal growth.

For me, it is a great pleasure to thank here those who who was important for me during the time in Rotterdam. Dear Lena, Masha, Sergey, Sasha, Julia, Marjana, Dragana, Robert, Vanessa, Aleksey, Nastia, Minghui, Diana, Marianna and many others: thanks a lot for the time we spent together doing interesting stuff. And especially, I would like to thank Yurii Moshkin: your remarkable sense of humor and endless topics for chats can make anyone feel cheerful (actually, during our chats I found out a lot about molecular biology from you).

And finally, Я хочу сказать спасибо самым важным для меня людям. Дорогая моя Лена, спасибо тебе за твою любовь и поддержку, в значительной степени благодаря которым была закончена эта книга. Дорогие мои Мама и Папа, я бесконечно обязан вам всем тем, что у меня есть. Именно те навыки и умения, которым вы меня научили, были ключевыми факторами для успешного преодоления всех преград. Ваша любовь, очевидно, была основной причиной появления этой книги. Спасибо.

7

CHAPTER: PHD PORTFOLIO SUMMARY

Name:	Maksim V Struchalin
Erasmus MC Department:	Epidemiology
Research School:	Netherlands Institute for Health Sciences (NIHES)
PhD period:	2007 - 2013
Promotors:	Prof.dr.ir. C.M. van Duijn and Prof.dr. B.A. Oostra
Co-promotor:	Dr. L.C. Karssen

	Year	ECTS
Courses		
Erasmus Summer Programmes	2007-2010	8.8
Study Design (CC01)	2010	4.3
Classical Methods for Data-analysis (CC02)	2009	5.7
Modern Statistical Methods (EP03)	2008	4.3
Genetic-Epidemiological Research Methods (GE02)	2008	5.7
SNPs and Human Diseases (GE08)	2008	1.4
Bayesian Statistics (CE09)	2008	0.9
Major Determinants and Major Diseases (EWP15)	2008	1.9
Advances in Population-based Studies of Complex Genetic Disorders (GE03)	2008	1.4
Genetic Linkage Analysis: Model Free Analysis (GE05)	2008	1.4
Presentations		
Annual presentations in Department of Epidemiology Rotterdam, Netherlands	2008-2012	5
Centre Nacional de Analisi Genomica, Bioinformatics Development Group, Barcelona, Spain	2011	1
Conferences and meetings		
6th annual CMSB meeting, Rotterdam, Netherlands	2009	1
CHARGE Consortium meeting, Rotterdam, Netherlands	2009	1
EMGM, University of Oxford, UK. Oral presentation.	2010	1
ENGAGE Consortium meeting, Barcelona, Spain. Oral presentation.	2010	1
ESHG Conference, Amsterdam, Netherlands	2011	1

PCDI PostDoc Retreat, Kapellerput, Netherlands	2012	1
Seminars and workshops		
Weekly scientific seminars in Department of Epidemiology, Rotterdam, Netherlands	2007-2012	1
GWAS, Invited workshop, Department of Neurology, Medical University of Graz, Austria	2009	1
Teaching activities		
GWAS Course, Erasmus Summer Program	2008-2010	1
Other		
Reviewer in BMC Bioinformatics		1
Reviewer in European Journal of Epidemiology		1

PUBLICATIONS

Publications with the first and second authorship

1. **Struchalin M**, Amin N, Eilers PH, van Duijn CM, Aulchenko YS, *An R package "VariABEL" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity*, BMC Genet. 2012 Jan 24;13:4.
2. Liu F, **Struchalin M**, Duijn K, Hofman A, Uitterlinden AG, Duijn C, Aulchenko YS, Kayser M, *Detecting low frequent loss-of-function alleles in genome wide association studies with red hair color as example*, PLoS One. 2011;6(11)
3. **Struchalin M**, Dehghan A, Witteman JC, van Duijn C, Aulchenko YS, *Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations*, BMC Genet. 2010 Oct 13;11:92.
4. Aulchenko YS, **Struchalin M**, van Duijn CM, *ProbABEL package for genome-wide association analysis of imputed data*, BMC Bioinformatics. 2010 Mar 16;11:134.
5. Aulchenko YS*, **Struchalin M***, Belonogova NM, Axenovich TI, Weedon MN, Hofman A, Uitterlinden AG, Kayser M, Oostra BA, van Duijn CM, Janssens AC, Borodin PM, *Predicting human height by Victorian and genomic methods*, Eur J Hum Genet. 2009 Aug;17(8):1070-5

Other publications with coauthorship

6. Köttgen A, Albrecht E, Teumer A, Vitart V, Krumsiek J, Hundertmark C, Pistis G, Ruggiero D, O'Seaghdha CM, Haller T, Yang Q, Tanaka T, Johnson AD, Kutalik Z, Smith AV, Shi J, **Struchalin M**, Middelberg RP, Brown MJ, Gaffo AL, Pirastu N, Li G, Hayward C, Zemunik T, Huffman J, Yengo L, Zhao JH, Demirkan A, Feitosa MF, Liu X, Malerba G, Lopez LM, van der Harst P, Li X, Kleber ME, Hicks AA, Nolte IM, Johansson A, Murgia F, Wild SH, Bakker SJ, Peden JF, Dehghan A, Steri M, Tenesa A, Lagou V, Salo P, Mangino M, Rose LM, Lehtimäki T, Woodward OM, Okada Y, Tin A, Müller C, Oldmeadow C, Putku M, Czamara D, Kraft P, Frogheri L, Thun GA, Grotevendt A, Gislason GK, Harris TB, Launer LJ, McArdle P, Shuldiner AR, Boerwinkle E, Coresh J, Schmidt H, Schallert M, Martin NG, Montgomery GW, Kubo M, Nakamura Y, Tanaka T, Munroe PB, Samani NJ, Jacobs DR Jr, Liu K, D'Adamo P, Ulivi S, Rotter JI, Psaty BM, Vollenweider P, Waeber G, Campbell S, Devuyst O, Navarro P, Kolcic I, Hastie N, Balkau B, Froguel P, Esko T, Salumets A, Khaw KT, Langenberg C, Wareham NJ, Isaacs A, Kraja A, Zhang Q, Wild PS, Scott RJ, Holliday EG, Org E, Viigimaa M, Bandinelli S, Metter JE, Lupo A, Trabetti E, Sorice R, Döring A, Lattka E, Strauch K, Theis F, Waldenberger M, Wichmann HE, Davies G, Gow AJ, Bruinenberg M; LifeLines Cohort Study, Stolk RP, Kooner JS, Zhang W, Winkelmann BR, Boehm BO, Lucae S, Penninx BW, Smit JH, Curhan G, Mudgal P, Plenge RM, Portas L, Persico I, Kirin M, Wilson JF, Mateo Leach I, van Gilst WH, Goel A, Ongen H, Hofman A, Rivadeneira F, Uitterlinden AG, Imboden M, von Eckardstein A, Cucca F, Nagaraja R, Piras MG, Nauck M, Schurmann C, Budde K, Ernst F, Farrington SM, Theodoratou E, Prokopenko I, Stumvoll M, Jula A, Perola M, Salomaa V, Shin SY, Spector TD, Sala C, Ridker PM, Kähönen M, Viikari J, Hengstenberg C, Nelson CP; CARDIoGRAM Consortium; DIAGRAM Consortium; ICBP Consortium; MAGIC Consortium, Meschia JF, Nalls MA, Sharma P, Singleton AB, Kamatani N, Zeller T, Burnier M, Attia J, Laan M, Klopp N, Hillege HL, Kloiber S, Choi H, Pirastu M, Tore S, Probst-Hensch NM, Völzke H, Gudnason V, Parsa A, Schmidt R, Whitfield

JB, Fornage M, Gasparini P, Siscovick DS, Polašek O, Campbell H, Rudan I, Bouatia-Naji N, Metspalu A, Loos RJ, van Duijn CM, Borecki IB, Ferrucci L, Gambaro G, Deary IJ, Wolffenbuttel BH, Chambers JC, März W, Pramstaller PP, Snieder H, Gyllensten U, Wright AF, Navis G, Watkins H, Witteman JC, Sanna S, Schipf S, Dunlop MG, Tönjes A, Ripatti S, Soranzo N, Toniolo D, Chasman DI, Raitakari O, Kao WH, Ciullo M, Fox CS, Caulfield M, Bochud M, Gieger C, *Genome-wide association analyses identify 18 new loci associated with serum urate concentrations*, Nat Genet. 2013 Feb;45(2):145-54.

7. Demirkan A, Isaacs A, Ugocsai P, Liebisch G, **Struchalin M**, Rudan I, Wilson JF, Pramstaller PP, Gyllensten U, Campbell H, Schmitz G, Oostra BA, van Duijn CM, *Plasma phosphatidylcholine and sphingomyelin concentrations are associated with depression and anxiety symptoms in a Dutch family-based lipidomics study*, J Psychiatr Res. 2012 Nov 30.
8. Chasman DI, Fuchsberger C, Pattaro C, Teumer A, Böger CA, Endlich K, Olden M, Chen MH, Tin A, Taliun D, Li M, Gao X, Gorski M, Yang Q, Hundertmark C, Foster MC, O'Seaghdha CM, Glazer N, Isaacs A, Liu CT, Smith AV, O'Connell JR, **Struchalin M**, Tanaka T, Li G, Johnson AD, Gierman HJ, Feitosa MF, Hwang SJ, Atkinson EJ, Lohman K, Cornelis MC, Johansson A, Tönjes A, Dehghan A, Lambert JC, Holliday EG, Sorice R, Kutalik Z, Lehtimäki T, Esko T, Deshmukh H, Ulivi S, Chu AY, Murgia F, Trompet S, Imboden M, Coassin S, Pistis G; CARDIoGRAM Consortium; ICBP Consortium; the CARe Consortium; WTCCC2, Harris TB, Launer LJ; Thor Aspelund, Eiriksdottir G, Mitchell BD, Boerwinkle E, Schmidt H, Cavalieri M, Rao M, Hu F, Demirkan A, Oostra BA, de Andrade M, Turner ST, Ding J, Andrews JS, Freedman BI, Giulianini F, Koenig W, Illig T, Meisinger C, Gieger C, Zgaga L, Zemunik T, Boban M, Minelli C, Wheeler HE, Igl W, Zaboli G, Wild SH, Wright AF, Campbell H, Ellinghaus D, Nöthlings U, Jacobs G, Biffar R, Ernst F, Homuth G, Kroemer HK, Nauck M, Stracke S, Völker U, Völzke H, Kovacs P, Stumvoll M, Mägi R, Hofman A, Uitterlinden AG, Rivadeneira F, Aulchenko YS, Polasek O, Hastie N, Vitart V, Helmer C, Wang JJ, Stengel B, Ruggiero D, Bergmann S, Kähönen M, Viikari J, Nikopensius T, Province M, Ketkar S, Colhoun H, Doney A, Robino A, Krämer BK, Portas L, Ford I, Buckley BM, Adam M, Thun GA, Paulweber B, Haun M, Sala C, Mitchell P, Ciullo M, Kim SK, Vollenweider P, Raitakari O, Metspalu A, Palmer C, Gasparini P, Pirastu M, Jukema JW, Probst-Hensch NM, Kronenberg F, Toniolo D, Gudnason V, Shuldiner AR, Coresh J, Schmidt R, Ferrucci L, Siscovick DS, van Duijn CM, Borecki IB, Kardia SL, Liu Y, Curhan GC, Rudan I, Gyllensten U, Wilson JF, Franke A, Pramstaller PP, Rettig R, Prokopenko I, Witteman J, Hayward C, Ridker PM, Parsa A, Bochud M, Heid IM, Kao WH, Fox CS, Kottgen A, *Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function*, Hum Mol Genet. 2012 Sep 25.
9. van Koolwijk LM, Ramdas WD, Ikram MK, Jansonius NM, Pasutto F, Hysi PG, Macgregor S, Janssen SF, Hewitt AW, Viswanathan AC, Ten Brink JB, Hosseini SM, Amin N, Despriet DD, Willemse-Assink JJ, Kramer R, Rivadeneira F, **Struchalin M**, Aulchenko YS, Weisschuh N, Zenkel M, Mardin CY, Gramer E, Welge-LüU, Montgomery GW, Carbonaro F, Young TL; The DCCT/EDIC Research Group, Bellenguez C, McGuffin P, Foster PJ, Topouzis F, Mitchell P, Wang JJ, Wong TY, Czudowska MA, Hofman A, Uitterlinden AG, Wolfs RC, de Jong PT, Oostra BA, Paterson AD; Wellcome Trust Case Control Consortium 2, Mackey DA, Bergen AA, Reis A, Hammond CJ, Vingerling JR, Lemij HG, Klaver CC, van Duijn CM., *Common Genetic Determinants of Intraocular Pressure and Primary Open-Angle Glaucoma*, PLoS Genet. 2012 May;8(5).
10. Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) Consortium, Bis JC, Decarli C, Smith AV, van der Lijn F, Crivello F, Fornage M, Debette S, Shulman JM, Schmidt H, Srikanth V, Schuur M, Yu L, Choi SH, Sigurdsson S, Verhaaren BF, Destefano AL, Lambert JC, Jack CR Jr, **Struchalin M**, Stankovich J, Ibrahim-Verbaas CA, Fleischman D, Zijdenbos A, den Heijer T, Mazoyer B, Coker LH, Enzinger C, Danoy P, Amin N, Arfanakis K, van Buchem MA, de Bruijn RF, Beiser A, Dufouil C, Huang J, Cavalieri M, Thomson R, Niessen WJ, Chibnik LB, Gislason GK, Hofman A, Pikula A, Amouyel P, Freeman KB, Phan TG, Oostra BA, Stein JL, Medland SE, Vasquez AA, Hibar DP, Wright MJ, Franke B, Martin NG, Thompson PM; the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, Nalls MA, Uitterlinden AG, Au R, Elbaz A, Beare RJ, van Swieten JC, Lopez OL, Harris TB, Chouraki V, Breteler MM, De Jager PL, Becker JT, Vernooij MW, Knopman D, Fazekas F, Wolf PA, van der

Lugt A, Gudnason V, Longstreth WT Jr, Brown MA, Bennett DA, van Duijn CM, Mosley TH, Schmidt R, Tzourio C, Launer LJ, Ikram MA, Seshadri S, *Common variants at 12q14 and 12q24 are associated with hippocampal volume*, Nat Genet. 2012 Apr 15.

11. the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, Ikram MA, Fornage M, Smith AV, Seshadri S, Schmidt R, Debette S, Vrooman HA, Sigurdsson S, Ropele S, Taal HR, Mook-Kanamori DO, Coker LH, Longstreth WT Jr, Niessen WJ, Destefano AL, Beiser A, Zijdenbos AP, **Struchalin M**, Jack CR Jr, Rivadeneira F, Uitterlinden AG, Knopman DS, Hartikainen AL, Pennell CE, Thiering E, Steegers EA, Hakonarson H, Heinrich J, Palmer LJ, Jarvelin MR, McCarthy MI, Grant SF, Pourcain BS, Timpson NJ, Smith GD, Sovio U, Nalls MA, Au R, Hofman A, Gudnason H, van der Lugt A, Harris TB, Meeks WM, Vernooij MW, van Buchem MA, Catellier D, Jaddoe VW, Gudnason V, Windham BG, Wolf PA, van Duijn CM, Mosley TH Jr, Schmidt H, Launer LJ, Breteler MM, Decarli C; Early Growth Genetics (EGG) Consortium, Adair LS, Ang W, Atalay M, van Beijsterveldt T, Bergen N, Benke K, Berry D, Coin L, Davis OS, Elliott P, Flexeder C, Frayling T, Gaillard R, Groen-Blokhuis M, Goh LK, Haworth CM, Hadley D, Hedebrand J, Hinney A, Hirschhorn JN, Holloway JW, Holst C, Jan Hottenga J, Horikoshi M, Huikari V, Hypponen E, Kilpeläen TO, Kirin M, Kowgier M, Lakka HM, Lange LA, Lawlor DA, Lehtimä T, Lewin A, Lindgren C, Lindi V, Maggi R, Marsh J, Middeldorp C, Millwood I, Murray JC, Nivard M, Nohr EA, Ntalla I, Oken E, Panoutsopoulou K, Pararajasingham J, Rodriguez A, Salem RM, Sebert S, Siitonen N, Strachan DP, Teo YY, Valcáel B, Willemsen G, Zeggini E, Boomsma DI, Cooper C, Gillman M, Hocher B, Lakka TA, Mohlke KL, Dedoussis GV, Ong KK, Pearson ER, Price TS, Power C, Raitakari OT, Saw SM, Scherag A, Simell O, Søen TI, Wilson JF, *Common variants at 6q22 and 17q21 are associated with intracranial volume*, Nat Genet. 2012 Apr 15.
12. Pattaro C, Köen A, Teumer A, Garnaas M, Bö CA, Fuchsberger C, Olden M, Chen MH, Tin A, Taliun D, Li M, Gao X, Gorski M, Yang Q, Hundertmark C, Foster MC, O'Seaghdha CM, Glazer N, Isaacs A, Liu CT, Smith AV, O'Connell JR, **Struchalin M**, Tanaka T, Li G, Johnson AD, Gierman HJ, Feitosa M, Hwang SJ, Atkinson EJ, Lohman K, Cornelis MC, Johansson A, Tös A, Dehghan A, Chouraki V, Holliday EG, Sorice R, Kutalik Z, Lehtimä T, Esko T, Deshmukh H, Ulivi S, Chu AY, Murgia F, Trompet S, Imboden M, Kollerits B, Pistis G; CARDIoGRAM Consortium; ICBP Consortium; CARe Consortium; Wellcome Trust Case Control Consortium 2 (WTCCC2), Harris TB, Launer LJ, Aspelund T, Eiriksdottir G, Mitchell BD, Boerwinkle E, Schmidt H, Cavalieri M, Rao M, Hu FB, Demirkan A, Oostra BA, de Andrade M, Turner ST, Ding J, Andrews JS, Freedman BI, Koenig W, Illig T, Dög A, Wichmann HE, Kolcic I, Zemunik T, Boban M, Minelli C, Wheeler HE, Igl W, Zaboli G, Wild SH, Wright AF, Campbell H, Ellinghaus D, Nöings U, Jacobs G, Biffar R, Endlich K, Ernst F, Homuth G, Kroemer HK, Nauck M, Stracke S, Vör U, Vöe H, Kovacs P, Stumvoll M, Mä R, Hofman A, Uitterlinden AG, Rivadeneira F, Aulchenko YS, Polasek O, Hastie N, Vitart V, Helmer C, Wang JJ, Ruggiero D, Bergmann S, Känen M, Viikari J, Nikopensius T, Province M, Ketkar S, Colhoun H, Doney A, Robino A, Giulianini F, Krär BK, Portas L, Ford I, Buckley BM, Adam M, Thun GA, Paulweber B, Haun M, Sala C, Metzger M, Mitchell P, Ciullo M, Kim SK, Vollenweider P, Raitakari O, Metspalu A, Palmer C, Gasparini P, Pirastu M, Jukema JW, Probst-Hensch NM, Kronenberg F, Toniolo D, Gudnason V, Shuldiner AR, Coresh J, Schmidt R, Ferrucci L, Siscovick DS, van Duijn CM, Borecki I, Kardia SL, Liu Y, Curhan GC, Rudan I, Gyllensten U, Wilson JF, Franke A, Pramstaller PP, Rettig R, Prokopenko I, Witteman JC, Hayward C, Ridker P, Parsa A, Bochud M, Heid IM, Goessling W, Chasman DI, Kao WH, Fox CS, *Genome-wide association and functional follow-up reveals new Loci for kidney function*, PLoS Genet. 2012 Mar;8(3).
13. Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G, Wilson JF, Johansson Å, Rudan I, Aulchenko YS, Kirichenko AV, Janssens AC, Jansen RC, Gnewuch C, Domingues FS, Pattaro C, Wild SH, Jonasson I, Polasek O, Zorkoltseva IV, Hofman A, Karssen LC, **Struchalin M**, Floyd J, Igl W, Biloglav Z, Broer L, Pfeufer A, Pichler I, Campbell S, Zaboli G, Kolcic I, Rivadeneira F, Huffman J, Hastie ND, Uitterlinden A, Franke L, Franklin CS, Vitart V; DIAGRAM Consortium, Nelson CP, Preuss M; CARDIoGRAM Consortium, Bis JC, O'Donnell CJ, Franceschini N; CHARGE Consortium,

Witteman JC, Axenovich T, Oostra BA, Meitinger T, Hicks AA, Hayward C, Wright AF, Gyllensten U, Campbell H, Schmitz G; EUROSPAN consortium, *Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations*, PLoS Genet. 2012 Feb;8(2).

14. Sanchez-Juan P, Bishop MT, Aulchenko YS, Brandel JP, Rivadeneira F, **Struchalin M**, Lambert JC, Amouyel P, Combarros O, Sainz J, Carracedo A, Uitterlinden AG, Hofman A, Zerr I, Kretzschmar HA, Laplanche JL, Knight RS, Will RG, van Duijn CM, *Genome-wide study links MTMR7 gene to variant Creutzfeldt-Jakob risk*, Neurobiol Aging. 2011 Nov 30.
15. Surakka I, Isaacs A, Karssen LC, Laurila PP, Middelberg RP, Tikkanen E, Ried JS, Lamina C, Mangino M, Igl W, Hottenga JJ, Lagou V, van der Harst P, Mateo Leach I, Esko T, Kutalik Z, Wainwright NW, **Struchalin M**, Sarin AP, Kangas AJ, Viikari JS, Perola M, Rantanen T, Petersen AK, Soininen P, Johansson A, Soranzo N, Heath AC, Papamarkou T, Prokopenko I, Tönjes A, Kronenberg F, Döring A, Rivadeneira F, Montgomery GW, Whitfield JB, Kähönen M, Lehtimäki T, Freimer NB, Willemsen G, de Geus EJ, Palotie A, Sandhu MS, Waterworth DM, Metspalu A, Stumvoll M, Uitterlinden AG, Jula A, Navis G, Wijmenga C, Wolffenbuttel BH, Taskinen MR, Ala-Korpela M, Kaprio J, Kyvik KO, Boomsma DI, Pedersen NL, Gyllensten U, Wilson JF, Rudan I, Campbell H, Pramstaller PP, Spector TD, Witteman JC, Eriksson JG, Salomaa V, Oostra BA, Raitakari OT, Wichmann HE, Gieger C, Järvelin MR, Martin NG, Hofman A, McCarthy MI, Peltonen L, van Duijn CM, Aulchenko YS, Ripatti S; ENGAGE Consortium, *A genome-wide screen for interactions reveals a new locus on 4p15 modifying the effect of waist-to-hip ratio on total cholesterol*, PLoS Genet. 2011 Oct;7(10).
16. Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, Ripatti S, Aulchenko YS, Zhang W, Yuan X, Lim N, Luan J, Ashford S, Wheeler E, Young EH, Hadley D, Thompson JR, Braund PS, Johnson T, **Struchalin M**, Surakka I, Luben R, Khaw KT, Rodwell SA, Loos RJ, Boekholdt SM, Inouye M, Deloukas P, Elliott P, Schlessinger D, Sanna S, Scuteri A, Jackson A, Mohlke KL, Tuomilehto J, Roberts R, Stewart A, Kesäniemi YA, Mahley RW, Grundy SM; Wellcome Trust Case Control Consortium, McArdle W, Cardon L, Waeber G, Vollenweider P, Chambers JC, Boehnke M, Abecasis GR, Salomaa V, Järvelin MR, Ruokonen A, Barroso I, Epstein SE, Hakonarson HH, Rader DJ, Reilly MP, Witteman JC, Hall AS, Samani NJ, Strachan DP, Barter P, van Duijn CM, Kooner JS, Peltonen L, Wareham NJ, McPherson R, Mooser V, Sandhu MS, *Genetic variants influencing circulating lipid levels and risk of coronary artery disease*, Arterioscler Thromb Vasc Biol. 2010 Nov;30(11):2264-76.
17. Solouki AM, Verhoeven VJ, van Duijn CM, Verkerk AJ, Ikram MK, Hysi PG, Despriet DD, van Koolwijk LM, Ho L, Ramdas WD, Czudowska M, Kuijpers RW, Amin N, **Struchalin M**, Aulchenko YS, van Rij G, Riemslag FC, Young TL, Mackey DA, Spector TD, Gorgels TG, Willemse-Assink JJ, Isaacs A, Kramer R, Swagemakers SM, Bergen AA, van Oosterhout AA, Oostra BA, Rivadeneira F, Uitterlinden AG, Hofman A, de Jong PT, Hammond CJ, Vingerling JR, Klaver CC, *A genome-wide association study identifies a susceptibility locus for refractive errors and myopia at 15q14*, Nat Genet. 2010 Oct;42(10):897-901.
18. Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV, O'Connell JR, Li M, Schmidt H, Tanaka T, Isaacs A, Ketkar S, Hwang SJ, Johnson AD, Dehghan A, Teumer A, Paré G, Atkinson EJ, Zeller T, Lohman K, Cornelis MC, Probst-Hensch NM, Kronenberg F, Tönjes A, Hayward C, Aspelund T, Eiriksdottir G, Launer LJ, Harris TB, Rampersaud E, Mitchell BD, Arking DE, Boerwinkle E, **Struchalin M**, Cavalieri M, Singleton A, Giallauria F, Metter J, de Boer IH, Haritunians T, Lumley T, Siscovick D, Psaty BM, Zillikens MC, Oostra BA, Feitosa M, Province M, de Andrade M, Turner ST, Schillert A, Ziegler A, Wild PS, Schnabel RB, Wilde S, Munzel TF, Leak TS, Illig T, Klopp N, Meisinger C, Wichmann HE, Koenig W, Zgaga L, Zemunik T, Kolcic I, Minelli C, Hu FB, Johansson A, Igl W, Zaboli G, Wild SH, Wright AF, Campbell H, Ellinghaus D, Schreiber S, Aulchenko YS, Felix JF, Rivadeneira F, Uitterlinden AG, Hofman A, Imboden M, Nitsch D, Brandstätter A, Kollerits B, Kedenko L, Mägi R, Stumvoll M, Kovacs P, Boban M, Campbell S, Endlich K, Völzke H, Kroemer HK, Nauck M, Völker U, Polasek O, Vitart V, Badola S, Parker AN, Ridker PM, Kardia SL, Blankenberg S, Liu Y, Curhan GC, Franke A, Rochat T, Paulweber B, Prokopenko I, Wang W, Gudnason V, Shuldiner AR, Coresh J, Schmidt R, Ferrucci L, Shlipak MG, van Duijn CM, Borecki I, Krämer BK, Rudan I, Gyllensten U, Wilson JF, Witteman JC,

Pramstaller PP, Rettig R, Hastie N, Chasman DI, Kao WH, Heid IM, Fox CS, *New loci associated with kidney function and chronic kidney disease*, Nat Genet. 2010 May;42(5):376-84.

19. Debette S, Bis JC, Fornage M, Schmidt H, Ikram MA, Sigurdsson S, Heiss G, **Struchalin M**, Smith AV, van der Lugt A, DeCarli C, Lumley T, Knopman DS, Enzinger C, Eiriksdottir G, Koudstaal PJ, DeStefano AL, Psaty BM, Dufouil C, Catellier DJ, Fazekas F, Aspelund T, Aulchenko YS, Beiser A, Rotter JI, Tzourio C, Shibata DK, Tscherner M, Harris TB, Rivadeneira F, Atwood LD, Rice K, Gottesman RF, van Buchem MA, Uitterlinden AG, Kelly-Hayes M, Cushman M, Zhu Y, Boerwinkle E, Gudnason V, Hofman A, Romero JR, Lopez O, van Duijn CM, Au R, Heckbert SR, Wolf PA, Mosley TH, Seshadri S, Breteler MM, Schmidt R, Launer LJ, Longstreth WT Jr, *Genome-wide association studies of MRI-defined brain infarcts: meta-analysis from the CHARGE Consortium*, Stroke. 2010 Feb;41(2):210-7.
20. Liu F, Ikram MA, Janssens AC, Schuur M, de Koning I, Isaacs A, **Struchalin M**, Uitterlinden AG, den Dunnen JT, Sleegers K, Bettens K, Van Broeckhoven C, van Swieten J, Hofman A, Oostra BA, Aulchenko YS, Breteler MM, van Duijn CM, *A study of the SORL1 gene in Alzheimer's disease and cognitive function*, J Alzheimers Dis. 2009;18(1):51-64.
21. Vasan RS, Glazer NL, Felix JF, Lieb W, Wild PS, Felix SB, Watzinger N, Larson MG, Smith NL, Dehghan A, Grosshennig A, Schillert A, Teumer A, Schmidt R, Kathiresan S, Lumley T, Aulchenko YS, König IR, Zeller T, Homuth G, **Struchalin M**, Aragam J, Bis JC, Rivadeneira F, Erdmann J, Schnabel RB, Dörr M, Zweiker R, Lind L, Rodeheffer RJ, Greiser KH, Levy D, Haritunians T, Deckers JW, Stritzke J, Lackner KJ, Völker U, Ingelsson E, Kullo I, Haerting J, O'Donnell CJ, Heckbert SR, Stricker BH, Ziegler A, Reffelmann T, Redfield MM, Werdan K, Mitchell GF, Rice K, Arnett DK, Hofman A, Gottdiener JS, Uitterlinden AG, Meitinger T, Blettner M, Friedrich N, Wang TJ, Psaty BM, van Duijn CM, Wichmann HE, Munzel TF, Kroemer HK, Benjamin EJ, Rotter JI, Witteman JC, Schunkert H, Schmidt H, Völzke H, Blankenberg S, *Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data*, JAMA. 2009 Jul 8;302(2):168-78.
22. Heard-Costa NL, Zillikens MC, Monda KL, Johansson A, Harris TB, Fu M, Haritunians T, Feitosa MF, Aspelund T, Eiriksdottir G, Garcia M, Launer LJ, Smith AV, Mitchell BD, McArdle PF, Shuldiner AR, Bielinski SJ, Boerwinkle E, Brancati F, Demerath EW, Pankow JS, Arnold AM, Chen YD, Glazer NL, McKnight B, Psaty BM, Rotter JI, Amin N, Campbell H, Gyllensten U, Pattaro C, Pramstaller PP, Rudan I, **Struchalin M**, Vitart V, Gao X, Kraja A, Province MA, Zhang Q, Atwood LD, Dupuis J, Hirschhorn JN, Jaquish CE, O'Donnell CJ, Vasan RS, White CC, Aulchenko YS, Estrada K, Hofman A, Rivadeneira F, Uitterlinden AG, Witteman JC, Oostra BA, Kaplan RC, Gudnason V, O'Connell JR, Borecki IB, van Duijn CM, Cupples LA, Fox CS, North KE, *NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium*, PLoS Genet. 2009 Jun;5(6).
23. Axenovich TI, Zorkoltseva IV, Belonogova NM, **Struchalin M**, Kirichenko AV, Kayser M, Oostra BA, van Duijn CM, Aulchenko YS, *Linkage analysis of adult height in a large pedigree from a Dutch genetically isolated population*, Hum Genet. 2009 Sep;126(3):457-71.
24. Ikram MA, Seshadri S, Bis JC, Fornage M, DeStefano AL, Aulchenko YS, Debette S, Lumley T, Folsom AR, van den Herik EG, Bos MJ, Beiser A, Cushman M, Launer LJ, Shahar E, **Struchalin M**, Du Y, Glazer NL, Rosamond WD, Rivadeneira F, Kelly-Hayes M, Lopez OL, Coresh J, Hofman A, DeCarli C, Heckbert SR, Koudstaal PJ, Yang Q, Smith NL, Kase CS, Rice K, Haritunians T, Roks G, de Kort PL, Taylor KD, de Lau LM, Oostra BA, Uitterlinden AG, Rotter JI, Boerwinkle E, Psaty BM, Mosley TH, van Duijn CM, Breteler MM, Longstreth WT Jr, Wolf PA, *Genomewide association studies of stroke*, N Engl J Med. 2009 Apr 23;360(17):1718-28.

25. Taal HR, St Pourcain B, Thiering E, Das S, Mook-Kanamori DO, Warrington NM, Kaakinen M, Kreiner-Møller E, Bradfield JP, Freathy RM, Geller F, Guxens M, Cousminer DL, Kerkhof M, Timpson NJ, Ikram MA, Beilin LJ, Bønnelykke K, Buxton JL, Charoen P, Chawes BL, Eriksson J, Evans DM, Hofman A, Kemp JP, Kim CE, Klopp N, Lahti J, Lye SJ, McMahon G, Mentch FD, Müller-Nurasyid M, O'Reilly PF, Prokopenko I, Rivadeneira F, Steegers EA, Sunyer J, Tiesler C, Yaghootkar H; Cohorts for Heart and Aging Research in Genetic Epidemiology Consortium, Breteler MM, Decarli C, Breteler MM, Debette S, Fornage M, Gudnason V, Launer LJ, van der Lugt A, Mosley TH Jr, Seshadri S, Smith AV, Vernooij MW; Early Genetics & Lifecourse Epidemiology Consortium, Blakemore AI, Chiavacci RM, Feenstra B, Fernandez-Banet J, Grant SF, Hartikainen AL, van der Heijden AJ, Iñiguez C, Lathrop M, McArdle WL, Mølgaard A, Newnham JP, Palmer LJ, Palotie A, Pouta A, Ring SM, Sovio U, Standl M, Uitterlinden AG, Wichmann HE, Vissing NH, DeCarli C, van Duijn CM, McCarthy MI, Koppelman GH, Estivill X, Hattersley AT, Melbye M, Bisgaard H, Pennell CE, Widen E, Hakonarson H, Smith GD, Heinrich J, Jarvelin MR, Jaddoe VW; Early Growth Genetics Consortium, *Common variants at 12q15 and 12q24 are associated with infant head circumference*, Nat Genet. 2012 Apr 15;44(5):532-8.
26. Stein JL, Medland SE, Vasquez AA, Hibar DP, Senstad RE, Winkler AM, Toro R, Appel K, Bartecek R, Bergmann Ø, Bernard M, Brown AA, Cannon DM, Chakravarty MM, Christoforou A, Domin M, Grimm O, Hollinshead M, Holmes AJ, Homuth G, Hottenga JJ, Langan C, Lopez LM, Hansell NK, Hwang KS, Kim S, Laje G, Lee PH, Liu X, Loth E, Lourdusamy A, Mattingsdal M, Mohnke S, Maniega SM, Nho K, Nugent AC, O'Brien C, Papmeyer M, Pütz B, Ramasamy A, Rasmussen J, Rijpkema M, Risacher SL, Roddey JC, Rose EJ, Ryten M, Shen L, Sprooten E, Strengman E, Teumer A, Trabzuni D, Turner J, van Eijk K, van Erp TG, van Tol MJ, Wittfeld K, Wolf C, Woudstra S, Aleman A, Alhusaini S, Almasy L, Binder EB, Brohawn DG, Cantor RM, Carless MA, Corvin A, Czisch M, Curran JE, Davies G, de Almeida MA, Delanty N, Depondt C, Duggirala R, Dyer TD, Erk S, Fagerness J, Fox PT, Freimer NB, Gill M, Göring HH, Hagler DJ, Hoehn D, Holsboer F, Hoogman M, Hosten N, Jahanshad N, Johnson MP, Kasperaviciute D, Kent JW Jr, Kochunov P, Lancaster JL, Lawrie SM, Liewald DC, Mandl R, Matarin M, Mattheisen M, Meisenzahl E, Melle I, Moses EK, Mühleisen TW, Nauck M, Nöthen MM, Olvera RL, Pandolfo M, Pike GB, Puls R, Reinvang I, Rentería ME, Rietschel M, Roffman JL, Royle NA, Rujescu D, Savitz J, Schnack HG, Schnell K, Seiferth N, Smith C, Steen VM, Valdés Hernández MC, Van den Heuvel M, van der Wee NJ, Van Haren NE, Veltman JA, Völzke H, Walker R, Westlye LT, Whelan CD, Agartz I, Boomsma DI, Cavalleri GL, Dale AM, Djurovic S, Drevets WC, Hagoort P, Hall J, Heinz A, Jack CR Jr, Foroud TM, Le Hellard S, Macciardi F, Montgomery GW, Poline JB, Porteous DJ, Sisodiya SM, Starr JM, Sussmann J, Toga AW, Veltman DJ, Walter H, Weiner MW; Alzheimer's Disease Neuroimaging Initiative; EPIGEN Consortium; IMAGEN Consortium; Saguenay Youth Study Group, Bis JC, Ikram MA, Smith AV, Gudnason V, Tzourio C, Vernooij MW, Launer LJ, DeCarli C, Seshadri S; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium, Andreassen OA, Apostolova LG, Bastin ME, Blangero J, Brunner HG, Buckner RL, Cichon S, Coppola G, de Zubicaray GI, Deary IJ, Donohoe G, de Geus EJ, Espeseth T, Fernández G, Glahn DC, Grabe HJ, Hardy J, Hulshoff Pol HE, Jenkinson M, Kahn RS, McDonald C, McIntosh AM, McMahon FJ, McMahon KL, Meyer-Lindenberg A, Morris DW, Müller-Myhsok B, Nichols TE, Ophoff RA, Paus T, Pausova Z, Penninx BW, Potkin SG, Sämann PG, Saykin AJ, Schumann G, Smoller JW, Wardlaw JM, Weale ME, Martin NG, Franke B, Wright MJ, Thompson PM; Enhancing Neuro Imaging Genetics through Meta-Analysis Consortium, *Identification of common variants associated with human hippocampal and intracranial volumes*, Nat Genet. 2012 Apr 15;44(5):552-61.
27. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Smith AV, Tobin MD, Verwoert GC, Hwang SJ, Pihur V, Vollenweider P, O'Reilly PF, Amin N, Bragg-Gresham JL, Teumer A, Glazer NL, Launer L, Zhao JH, Aulchenko Y, Heath S, Sõber S, Parsa A, Luan J, Arora P, Dehghan A, Zhang F, Lucas G, Hicks AA, Jackson AU, Peden JF, Tanaka

T, Wild SH, Rudan I, Igl W, Milaneschi Y, Parker AN, Fava C, Chambers JC, Fox ER, Kumari M, Go MJ, van der Harst P, Kao WH, Sjögren M, Vinay DG, Alexander M, Tabara Y, Shaw-Hawkins S, Whincup PH, Liu Y, Shi G, Kuusisto J, Tayo B, Seielstad M, Sim X, Nguyen KD, Lehtimäki T, Matullo G, Wu Y, Gaunt TR, Onland-Moret NC, Cooper MN, Platou CG, Org E, Hardy R, Dahgam S, Palmen J, Vitart V, Braund PS, Kuznetsova T, Uiterwaal CS, Adeyemo A, Palmas W, Campbell H, Ludwig B, Tomaszewski M, Tzoulaki I, Palmer ND; CARDIoGRAM consortium; CKDGen Consortium; KidneyGen Consortium; EchoGen consortium; CHARGE-HF consortium, Aspelund T, Garcia M, Chang YP, O'Connell JR, Steinle NI, Grobbee DE, Arking DE, Kardia SL, Morrison AC, Hernandez D, Najjar S, McArdle WL, Hadley D, Brown MJ, Connell JM, Hingorani AD, Day IN, Lawlor DA, Beilby JP, Lawrence RW, Clarke R, Hopewell JC, Ongen H, Dreisbach AW, Li Y, Young JH, Bis JC, Kähönen M, Viikari J, Adair LS, Lee NR, Chen MH, Olden M, Pattaro C, Bolton JA, Köttgen A, Bergmann S, Mooser V, Chaturvedi N, Frayling TM, Islam M, Jafar TH, Erdmann J, Kulkarni SR, Bornstein SR, Grässler J, Groop L, Voight BF, Kettunen J, Howard P, Taylor A, Guarrera S, Ricceri F, Emilsson V, Plump A, Barroso I, Khaw KT, Weder AB, Hunt SC, Sun YV, Bergman RN, Collins FS, Bonnycastle LL, Scott LJ, Stringham HM, Peltonen L, Perola M, Vartiainen E, Brand SM, Staessen JA, Wang TJ, Burton PR, Soler Artigas M, Dong Y, Snieder H, Wang X, Zhu H, Lohman KK, Rudock ME, Heckbert SR, Smith NL, Wiggins KL, Doumatey A, Shriner D, Veldre G, Viigimaa M, Kinra S, Prabhakaran D, Tripathy V, Langefeld CD, Rosengren A, Thelle DS, Corsi AM, Singleton A, Forrester T, Hilton G, McKenzie CA, Salako T, Iwai N, Kita Y, Ogihara T, Ohkubo T, Okamura T, Ueshima H, Umemura S, Eyheramendy S, Meitinger T, Wichmann HE, Cho YS, Kim HL, Lee JY, Scott J, Sehmi JS, Zhang W, Hedblad B, Nilsson P, Smith GD, Wong A, Narisu N, Stančáková A, Raffel LJ, Yao J, Kathiresan S, O'Donnell CJ, Schwartz SM, Ikram MA, Longstreth WT Jr, Mosley TH, Seshadri S, Shrine NR, Wain LV, Morken MA, Swift AJ, Laitinen J, Prokopenko I, Zitting P, Cooper JA, Humphries SE, Danesh J, Rasheed A, Goel A, Hamsten A, Watkins H, Bakker SJ, van Gilst WH, Janipalli CS, Mani KR, Yajnik CS, Hofman A, Mattace-Raso FU, Oostra BA, Demirkan A, Isaacs A, Rivadeneira F, Lakatta EG, Orru M, Scuteri A, Ala-Korpela M, Kangas AJ, Lyytikäinen LP, Soininen P, Tukiainen T, Würtz P, Ong RT, Dörr M, Kroemer HK, Völker U, Völzke H, Galan P, Hercberg S, Lathrop M, Zelenika D, Deloukas P, Mangino M, Spector TD, Zhai G, Meschia JF, Nalls MA, Sharma P, Terzic J, Kumar MV, Denniff M, Zukowska-Szczechowska E, Wagenknecht LE, Fowkes FG, Charchar FJ, Schwarz PE, Hayward C, Guo X, Rotimi C, Bots ML, Brand E, Samani NJ, Polasek O, Talmud PJ, Nyberg F, Kuh D, Laan M, Hveem K, Palmer LJ, van der Schouw YT, Casas JP, Mohlke KL, Vineis P, Raitakari O, Ganesh SK, Wong TY, Tai ES, Cooper RS, Laakso M, Rao DC, Harris TB, Morris RW, Dominiczak AF, Kivimaki M, Marmot MG, Miki T, Saleheen D, Chandak GR, Coresh J, Navis G, Salomaa V, Han BG, Zhu X, Kooner JS, Melander O, Ridker PM, Bandinelli S, Gyllensten UB, Wright AF, Wilson JF, Ferrucci L, Farrall M, Tuomilehto J, Pramstaller PP, Elosua R, Soranzo N, Sijbrands EJ, Altshuler D, Loos RJ, Shuldiner AR, Gieger C, Meneton P, Uitterlinden AG, Wareham NJ, Gudnason V, Rotter JI, Rettig R, Uda M, Strachan DP, Witteman JC, Hartikainen AL, Beckmann JS, Boerwinkle E, Vasan RS, Boehnke M, Larson MG, Järvelin MR, Psaty BM, Abecasis GR, Chakravarti A, Elliott P, van Duijn CM, Newton-Cheh C, Levy D, Caulfield MJ, Johnson T, *Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk*, Nature. 2011 Sep 11;478(7367):103-9.

28. Wain LV, Verwoert GC, O'Reilly PF, Shi G, Johnson T, Johnson AD, Bochud M, Rice KM, Henneman P, Smith AV, Ehret GB, Amin N, Larson MG, Mooser V, Hadley D, Dörr M, Bis JC, Aspelund T, Esko T, Janssens AC, Zhao JH, Heath S, Laan M, Fu J, Pistis G, Luan J, Arora P, Lucas G, Pirastu N, Pichler I, Jackson AU, Webster RJ, Zhang F, Peden JF, Schmidt H, Tanaka T, Campbell H, Igl W, Milaneschi Y, Hottenga JJ, Vitart V, Chasman DI, Trompet S, Bragg-Gresham JL, Alizadeh BZ, Chambers JC, Guo X, Lehtimäki T, Kühnel B, Lopez LM, Polašek O, Boban M, Nelson CP, Morrison AC, Pihur V, Ganesh SK, Hofman A, Kundu S, Mattace-Raso FU, Rivadeneira F, Sijbrands EJ, Uitterlinden AG, Hwang SJ, Vasan RS, Wang TJ, Bergmann S, Vollenweider P, Waeber G, Laitinen J, Pouta A, Zitting P, McArdle WL, Kroemer HK, Völker U, Völzke H, Glazer NL, Taylor KD, Harris TB, Alavere H, Haller T, Keis A, Tammesoo ML, Aulchenko Y, Barroso I, Khaw KT, Galan P, Hercberg S, Lathrop M, Eyheramendy S, Org E, Sõber S, Lu X, Nolte IM, Penninx BW, Corre T, Masciullo C, Sala C, Groop L, Voight BF, Melander O, O'Donnell CJ, Salomaa V, d'Adamo AP, Fabretto A, Faletra F, Ulivi

S, Del Greco F, Facheris M, Collins FS, Bergman RN, Beilby JP, Hung J, Musk AW, Mangino M, Shin SY, Soranzo N, Watkins H, Goel A, Hamsten A, Gider P, Loitfelder M, Zeginigg M, Hernandez D, Najjar SS, Navarro P, Wild SH, Corsi AM, Singleton A, de Geus EJ, Willemsen G, Parker AN, Rose LM, Buckley B, Stott D, Orru M, Uda M; LifeLines Cohort Study, van der Klauw MM, Zhang W, Li X, Scott J, Chen YD, Burke GL, Kähönen M, Viikari J, Döring A, Meitinger T, Davies G, Starr JM, Emilsson V, Plump A, Lindeman JH, Hoen PA, König IR; EchoGen consortium, Felix JF, Clarke R, Hopewell JC, Ongen H, Breteler M, Debette S, Destefano AL, Fornage M; AortaGen Consortium, Mitchell GF; CHARGE Consortium Heart Failure Working Group, Smith NL; KidneyGen consortium, Holm H, Stefansson K, Thorleifsson G, Thorsteinsdottir U; CKDGen consortium; Cardiogenics consortium; CardioGram, Samani NJ, Preuss M, Rudan I, Hayward C, Deary IJ, Wichmann HE, Raitakari OT, Palmas W, Kooner JS, Stolk RP, Jukema JW, Wright AF, Boomsma DI, Bandinelli S, Gyllensten UB, Wilson JF, Ferrucci L, Schmidt R, Farrall M, Spector TD, Palmer LJ, Tuomilehto J, Pfeufer A, Gasparini P, Siscovick D, Altshuler D, Loos RJ, Toniolo D, Snieder H, Gieger C, Meneton P, Wareham NJ, Oostra BA, Metspalu A, Launer L, Rettig R, Strachan DP, Beckmann JS, Witteman JC, Erdmann J, van Dijk KW, Boerwinkle E, Boehnke M, Ridker PM, Jarvelin MR, Chakravarti A, Abecasis GR, Gudnason V, Newton-Cheh C, Levy D, Munroe PB, Psaty BM, Caulfield MJ, Rao DC, Tobin MD, Elliott P, van Duijn CM, *Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure*, Nat Genet. 2011 Sep 11;43(10):1005-11.

29. Böger CA, Chen MH, Tin A, Olden M, Köttgen A, de Boer IH, Fuchsberger C, O'Seaghdha CM, Pattaro C, Teumer A, Liu CT, Glazer NL, Li M, O'Connell JR, Tanaka T, Peralta CA, Kutalik Z, Luan J, Zhao JH, Hwang SJ, Akylbekova E, Kramer H, van der Harst P, Smith AV, Lohman K, de Andrade M, Hayward C, Kollerits B, Tönjes A, Aspelund T, Ingelsson E, Eiriksdottir G, Launer LJ, Harris TB, Shuldiner AR, Mitchell BD, Arking DE, Franceschini N, Boerwinkle E, Egan J, Hernandez D, Reilly M, Townsend RR, Lumley T, Siscovick DS, Psaty BM, Kestenbaum B, Haritunians T, Bergmann S, Vollenweider P, Waeber G, Mooser V, Waterworth D, Johnson AD, Florez JC, Meigs JB, Lu X, Turner ST, Atkinson EJ, Leak TS, Aasarød K, Skorpen F, Syvänen AC, Illig T, Baumert J, Koenig W, Krämer BK, Devuyst O, Mychaleckyj JC, Minelli C, Bakker SJ, Kedenko L, Paulweber B, Coassin S, Endlich K, Kroemer HK, Biffar R, Stracke S, Völzke H, Stumvoll M, Mägi R, Campbell H, Vitart V, Hastie ND, Gudnason V, Kardia SL, Liu Y, Polasek O, Curhan G, Kronenberg F, Prokopenko I, Rudan I, Arnlöv J, Hallan S, Navis G; CKDGen Consortium, Parsa A, Ferrucci L, Coresh J, Shlipak MG, Bull SB, Paterson NJ, Wichmann HE, Wareham NJ, Loos RJ, Rotter JI, Pramstaller PP, Cupples LA, Beckmann JS, Yang Q, Heid IM, Rettig R, Dreisbach AW, Bochud M, Fox CS, Kao WH, *CUBN is a gene locus for albuminuria*, J Am Soc Nephrol. 2011 Mar;22(3):555-70.

30. Okada Y, Sim X, Go MJ, Wu JY, Gu D, Takeuchi F, Takahashi A, Maeda S, Tsunoda T, Chen P, Lim SC, Wong TY, Liu J, Young TL, Aung T, Seielstad M, Teo YY, Kim YJ, Lee JY, Han BG, Kang D, Chen CH, Tsai FJ, Chang LC, Fann SJ, Mei H, Rao DC, Hixson JE, Chen S, Katsuya T, Isono M, Ogihara T, Chambers JC, Zhang W, Kooner JS; KidneyGen Consortium; CKDGen Consortium, Albrecht E; GUGC consortium, Yamamoto K, Kubo M, Nakamura Y, Kamatani N, Kato N, He J, Chen YT, Cho YS, Tai ES, Tanaka T, *Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations*, Nat Genet. 2012 Jul 15;44(8):904-9.

** Those authors contributed equally*

ABOUT THE AUTHOR

Maksim Struchalin was born in Academgorodok, Novosibirsk, Russia in 1982. After graduating from high school on 1999, he was admitted to Novosibirsk State Technical University on the Physical-Technical Faculty where he studied physics of the nucleus and elementary particles. Two years later he was part-time employed as a laboratory assistant in laboratory 3-2, Budker's Institute of Nuclear Physics (BINP), Russia. Later, a scientific project which he performed at BINP was recognized as the best student project presented on the student conference "The days of Science NSTU-2005" in the section "physics" and got funding support from Novosibirsk State Technical University. In 2005, during his Master study and scientific work in BINP, Maksim was part-time employed as a scientific programmer in the Laboratory of Genetic Recombination and Segregation, Novosibirsk Institute of Cytology and Genetics, Russia where he worked for a year. In 2007, after a year of PhD study in BINP, he moved to Rotterdam, The Netherlands where he obtained a Master of Science degree in epidemiology and started a PhD study in the Epidemiology Department, Erasmus MC University Medical Center (head: Prof.dr. A. Hofman) in the genetic epidemiology unit (head: Prof.dr.ir. C.M. van Duijn). During his PhD study, Maksim co-authored about thirty publications, was involved in teaching activities and played an important role in bringing innovative computational hardware to the ErasmusMC in the framework of the Academic Partnership Program of nVidia Corporation