Statistical applications in nutrigenomics

Analyzing multiple genes and proteins
in relation to
complex diseases in humans

Promotoren:	Prof. dr. ir. E.J.M. Feskens Persoonlijk hoogleraar bij de afdeling Humane voeding Wageningen Universiteit
	Prof. dr. E.C.M. Mariman Hoogleraar Functionele genetica Universiteit Maastricht
Co-promotor:	Dr. ir. J.M.A. Boer Senior onderzoeker RIVM, Bilthoven
Promotiecommissie:	Prof. dr. ir. J. Keijer (Wageningen Universiteit) Dr. C. Lai (Tufts University, Boston, USA) Prof. dr. P.E. Slagboom (LUMC, Leiden) Prof. dr. A.H. Zwinderman (AMC, Amsterdam)

Dit onderzoek is uitgevoerd binnen de onderzoekschool VLAG

Statistical applications in nutrigenomics

Analyzing multiple genes and proteins in relation to complex diseases in humans

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof. dr. M.J. Kropff,
in het openbaar te verdedigen
op dinsdag 9 december 2008
des namiddags te half twee in de Aula

Geert Heidema
Statistical applications in nutrigenomics: analyzing multiple genes and proteins in relation to complex diseases in humans
Thesis Wageningen University, Wageningen, The Netherlands, 2008
ISBN: 978-90-858-261-2

Abstract

Background

The recent advances in technology provide the possibility to obtain large genomic datasets that contain information on large numbers of variables, while the sample sizes are moderate to small. This has lead to statistical challenges in the analysis of multiple genes and proteins in relation to complex diseases. In this thesis approaches are investigated to analyze large genomic datasets, taking complex relationships between genes, proteins and complex diseases into account. These approaches are applied to real data to investigate whether biologically relevant information from the dataset could be obtained or whether models could be obtained that are useful for diagnostic or prognostic purposes.

Results

We developed a general framework for the analysis of genetic, transcriptomic and proteomic data to obtain insight in biological mechanisms. This framework consists of the following steps: detection of heterogeneity, dimensionality reduction to deal with the large numbers of variables, statistical interpretation and biological interpretation. We found that within this multi-step approach application of a combination of methods, including methods that take interactions into account, is useful within the dimensionality reduction step. In this way more information is captured compared to applying only one method. After selection of relevant variables in the dimensionality reduction step, applying visualization tools, e.g. the interaction entropy graph, together with traditional statistical methods showed to be helpful for statistical interpretation whether variables contribute by their main and/or interaction effect to the outcome of interest. In the last step, biological interpretation of the statistical results was facilitated by literature search, pathway analysis and database mining.

Discussion

The general framework discussed in this thesis provides the possibility to analyze large nutrigenomic datasets. Although the contribution of genomic research to public health is at the moment limited, new advances in genomic research, e.g. genome-wide association studies, statistical approaches as discussed in this thesis, are promising and genomic research might in the near future lead to applications that translate into improvement of public health.

Contents

Chapter 1

Introduction

Background

Diagnosis, prognosis and insight in biological mechanisms

Approaches to analyze large genomic datasets to obtain insight in biological mechanisms

Simulation studies and applied research

Outline of the thesis

Background

Proteins carry out the biological processes in the human body. The proper levels and functioning of proteins are therefore of vital importance for the health and well-being of humans. The synthesis of proteins is encoded by genes. Genes are transcribed into RNA, and after post-transcriptional modification mRNA is formed. The sequence of the mRNA determines the sequence of the amino acids, and thereby the polypeptide that is synthesized at the ribosomes. Finally, post-translational modification of the polypeptide results in the formation of the active protein. Thus, the synthesis of proteins is determined by the genetic template of the DNA and the flow of information is from the DNA to the RNA to the protein [1].

The biological information at the levels of DNA, mRNA and proteins is studied by genetics, transcriptomics and proteomics, respectively. These research areas are included within the broader field of genomics. Two examples of the application of genomics are nutrigenomics and pharmacogenomics. Within the field of nutrigenomics, the aim is to identify interactions between genes and nutrition that are involved in health and disease. Research is performed to investigate in what way nutrition affects the expression of genes and how the genetic background of the individual responds to nutrients and bioactives [2]. In pharmacogenomics, instead of nutrition, the response to medication depending on the genetic background of the individual is studied.

Nowadays, due to the fast development in technology, data on huge numbers of genetic polymorphisms, mRNA levels, peptide and protein concentrations can be obtained by chips, microarrays, mass-spectrometry and multiplex assays, respectively. This allows researchers to obtain not only information about individual effects of genes and proteins, but also to study entire processes that are involved in the development of complex diseases. In the development of complex diseases, such as cardiovascular diseases, diabetes and different types of cancer, large numbers of factors are likely to be involved, each contributing with a moderate to small effect. Furthermore, these factors are assumed to be associated with the complex disease by ubiquitous and intricate interactions. Providing the possibility to measure large numbers of variables, the technological platforms enable researchers to study complex diseases in a more global way. However, researchers are faced with a huge amount of data and the challenge is to extract information from the data that is both relevant and interpretable. The moderate to small effects, together with the complexity of many interactions, increases the difficulty to detect the biologically relevant variables and the mechanisms that underlie the development of the disease. To test the influence of large numbers of variables and their possible interactions to one or several outcomes is not straightforward and has given rise to challenges in the statistical analyses of nutrigenomic data. The major statistical issues that are encountered in the analyses of genomic datasets include the dimensionality problem [3] and the multiple testing problem. Another statistical issue concerns the possible presence of heterogeneity [4]. These statistical challenges are discussed in more detail in chapter 2. In an attempt to overcome these statistical issues, methods have recently been developed for analyses of large numbers of single nucleotide polymorphisms (SNPs), mRNA levels and/or proteins, and more approaches to study huge amounts of data are currently under development.

Diagnosis, prognosis and insight in biological mechanisms

In genomic studies investigating complex diseases the interest is to obtain models for diagnosis, prognosis and/or to obtain insight in the biological mechanisms. In diagnosis, distinguishing between diseased individuals (cases) and healthy individuals (controls) is of importance. With prognosis the objective is to find biomarkers with prognostic value to detect, among others, at a preclinical stage the risk for disease. Statistical analyses for diagnostic purposes aim to obtain models that have a good performance in correctly classifying cases and controls. Analyses for prognosis aim to obtain models that correctly indicate for apparently healthy subjects the risk on developing a certain disease in the future. 'Black box' approaches that do not provide insight in the model itself but do perform well can be applied, such as Support Vector Machines (SVM) [5]. However, even for diagnosis and prognosis, interpretability of the model might be useful for and preferred by the clinician. In classification of disease status, the aim is to find the smallest set of variables that still has equal performance to larger sets of variables.

On the other hand, to obtain insight in the biological mechanisms, the largest possible amount of information needs to be retrieved from the data. To obtain this information it is important to discern between true signals and noise. However, this distinction is not always clear-cut, but may be a gray zone. This is especially the case in nutrigenomic studies, in which small effects that contribute by their interactions can be expected. With large genomic datasets, including information on genetic polymorphisms, mRNA levels, or protein concentrations, the consideration of the threshold to be used to select variables is to some extent subjective. However, the threshold to be applied to reduce noise should lead to biologically relevant, but also practical and interpretable models. From a holistic point of view, all variables are involved to some extent. However, we need to select a subset of important variables involved that can be worked with in practice. Therefore, methods that yield clear interpretable models need to be applied.

Approaches to analyze large genomic datasets to obtain insight in biological mechanisms

Different approaches can be taken to analyze large genomic datasets to understand complex traits and diseases. Statistical applications on the one hand are needed to provide insight in the significance of associations between variables and the outcome, biological information on the other hand is needed to verify whether results obtained from statistical analyses do have biological meaning. One approach is to start from statistics and move towards biological interpretation, applying statistical methods to select important variables and interpret how these variables are related to the outcome of interest. A second approach would be to incorporate biological knowledge in the statistical analyses. For example, variables can be grouped based on biological similarity, or their presence in the same biological process, and statistical analyses are subsequently performed to examine whether these groups significantly contribute to the endpoint of interest. In microarray data, gene set enrichment analysis (GSEA) [6] is an example of the second approach.

An example of the first approach will now be discussed in more detail. This multi-step approach (see figure 1) consists of the following points:

- Detection of heterogeneity
- Dimensionality reduction
- Statistical interpretation
- Biological interpretation

This general multi-step approach can be applied to the different biological levels, DNA, mRNA and proteins, to identify SNPs, genes and proteins, respectively, and interpret their relationship with complex diseases. Although these processes are described as subsequent steps, in practice applying statistical methods and interpreting the biological relevance may be an interactive process. The different steps will now be discussed in more detail.

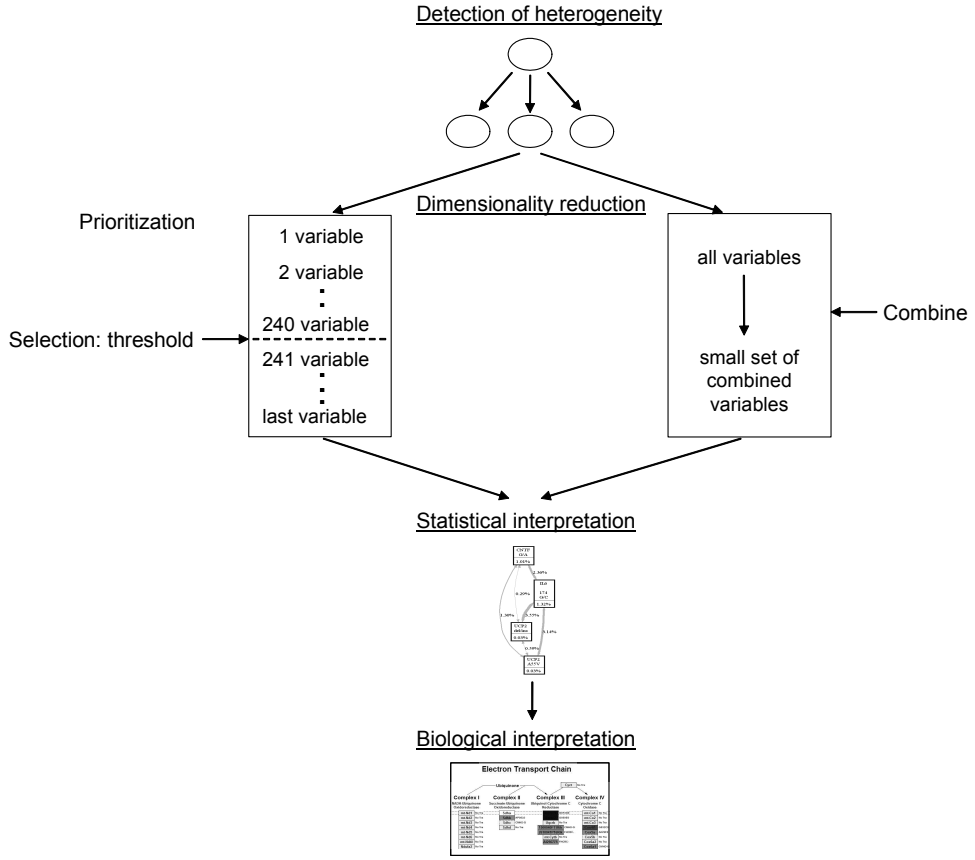


Figure 1: Multi-step approach to analyze large genomic datasets. Within the dimensionality reduction step two approaches are shown. On the left side the number of variables is reduced by prioritizing all variables and selecting a subset of most important variables by a threshold. On the right side the number of variables is reduced by combining variables into a smaller set of new variables.

Detection of heterogeneity

In genomic studies, intervention studies such as randomized controlled trials are underpowered if heterogeneity is present, as differences in responses to the treatment in the patient group due to individual background differences may affect the findings [7]. Subsequently, efficiency of interventions will be increased if directed to the appropriate subgroups. The presence of heterogeneity will influence the results if the statistical analyses are performed over the total population. Therefore, detection of heterogeneity is important as a first step before performing the planned analyses. To be able to detect heterogeneity, extensive information on covariate data (including potential genetic and environmental risk factors and phenotypic data) for the study population should be present [4].

Dimensionality reduction

Different approaches to reduce dimensionality exist (see figure 1). One approach is to prioritize the total number of variables in the dataset and subsequently apply a threshold to select a subset of important variables for further statistical and biological interpretation. In the prioritization step, variables are ranked by a certain measure of importance, for example the χ^2 or t-statistic. However, univariate test-statistics do not take possible effects of combinations of variables into account. Capturing all possible interactions between variables in the prioritization step is important, especially in nutrigenomic studies investigating complex diseases, in which many variables with small effects mainly contribute in interaction with each other. For this purpose multivariate statistical tools are needed.

The critical point in the step of selecting variables is how to define the threshold to select a subset of important variables, as it influences further interpretation. Defining the threshold depends on the type of study, which can have different signal to noise ratios, i.e. the levels of signals relative to the level of the background noise. For example, signals are normally much stronger in pharmacogenomic studies compared to nutrigenomic studies. In general, too stringent thresholds results in selection of the most important effects and very few false positive results, but leaves out many true positive results that contribute with weak effects. This is especially the case in nutritional intervention studies, in which the diet may result in many weak effects that by themselves are not strong enough, but in combination may result in a strong effect [8]. On the other hand, if the threshold is defined too liberal, many false positive results will be included and subsequent biological interpretations can be flawed. Thus, defining the threshold is not always straightforward, and is a balance between signal and noise, or in other words true and false positive results.

A second approach to reduce the dimensionality is to apply methods that combine variables into a smaller set of new variables. These newly created variables are subsequently related to the endpoint of interest. For example, partial least squares (PLS) [9, 10] reduces a set of continuous variables to a small number of latent components, taking the covariance structure with another (set of) outcome variable(s) into account. The multifactor dimensionality reduction (MDR) method [11-13] is a method to study interactions between SNPs, which reduces the dimensionality for combinations of SNPs by constructing a new variable. This new variable is subsequently used for classification of the endpoint [13].

Statistical interpretation

Statistical interpretation is useful to understand how the selected variables are related to the disease of interest, whether they contribute by their main effect and/or in interaction with other variables. Statistical interpretation can also provide insight in the direction and the strength of the association between variables and the disease. Visualization of effects is a practical means to interpret the results of the statistical analyses. Visualization tools, e.g. interaction graphs [13-15], are available that show graphically the importance of individual variables as well as the interactions between variables.

Biological interpretation

The biological interpretation of the statistical results is an important last step to verify whether the statistical results are biologically meaningful. If the results have not been reported previously in the literature, the plausibility of the results should be verified from a biological perspective. Also, the statistical results need to be confirmed by mechanistic studies, for example real-time PCR to confirm whether selected genes from microarray data are truly differentially expressed.

Statistical interaction found in the previous step does not per definition imply that biological interaction between the interacting variables is present: statistical interaction is measured on the population level whereas biological interaction is the physical interaction between molecules at the individual level [16]. For example, biological compounds may statistically interact in relation to disease, but not to interact at the individual biological level; biological compounds can be present within the same pathway but work together via different molecules, or the compounds can be present in different pathways and in that way may together contribute to the development of disease. Therefore, statistical interaction at the population level should be followed by mechanistic studies to gain more knowledge about the underlying biological mechanism that leads to the development of disease.

Simulation studies and applied research

To obtain insight in the biological mechanisms, it is important to know whether statistical methods do extract the relevant biological information from the data at hand. As in complex diseases many interactions are likely to play a role, it is of interest to know whether a method has the power to detect these interactions. But also other features are of importance, such as the ability to detect the presence of heterogeneity. Simulation studies to measure the performance of methods provides insight in their applicability; for example, they may show that a method is able to detect interactions under certain conditions, whereas another method may be suitable to detect important effects while controlling the type I error very well. These studies provide useful information about what methods are suitable to apply for the purpose of the study at hand, as well as useful information for the interpretability of their results. The insight simulation studies provide, can also be used to choose a combination of methods with different strengths and weaknesses for analyzing real data, covering better the detection of different types of effects. Programs are available for the generation of simulated datasets for e.g. SNPs [17, 18], microarray data [19] and mass spectrometry data [20].

On the other hand, simulation studies are based on synthetic data generated by the biostatistician. To mimic data obtained from a real-life study may appear to be difficult. For example, mimicking in microarray data the presence of complex interactions between genes involved in a dietary intervention study is a daunting task. Therefore, besides performing simulation studies to test the performance of methods, application of these methods to real-life data that show their ability to retrieve biologically relevant information from the data is also of importance. Thus, to test the applicability of methods, simulation studies need to go hand in hand with applications to real data.

Outline of the thesis

The objective of the research performed in this thesis was to study what methods are available to analyze large numbers of variables (e.g. SNPs, genes, proteins) in moderate to small sample sizes, overcoming the statistical problems encountered in the analysis of nutrigenomic data. Methods to analyze this type of data are described for both categorical and continuous endpoints. Furthermore, approaches to analyze nutrigenomic data have been investigated for the purposes of diagnosis, prognosis and obtaining insight in the biological mechanisms. Applications of these approaches to real data are shown for SNPs, mRNA levels, peptide masses and protein concentrations. Endpoints that were studied include coronary heart disease (CHD), body mass index (BMI), weight gain, high-density lipoprotein cholesterol (HDL-C), total cholesterol (total-C) and breast cancer.

This thesis includes chapters on statistical analyses in the field of genetic epidemiology to study SNP data (chapters 2-4), continues with the analyses of microarray data (chapter 5), proteomic data (chapter 6 and 7) and finally the results presented in this thesis are placed in a broader perspective in the general discussion. A more detailed outline of these chapters is described below.

Genetic epidemiology

An overview of several available methods to analyze associations of SNPs with categorical or continuous endpoints is discussed in chapter 2. These methods are referred to as multi-locus methods. Chapter 3 describes the application of different multi-locus methods to a real dataset to compare their results regarding the prioritization and selection of SNPs. In this chapter, after prioritization and selection of SNPs, application of two types of interaction graphs for statistical interpretation of how these SNPs contribute to disease is subsequently shown. In chapter 4, a combination of the interaction entropy graph and logistic regression analysis for statistical interpretation of a biological model of weight regulation is presented.

Transcriptomics

We investigated in chapter 5 the application of the multivariate method random forests (RF) to two real microarray datasets to take interactions into account in the selection of genes, and compare RF with the conventional t-test. Genes selected by RF were subsequently analyzed by self-organizing maps (SOM) to find clusters containing genes with similar gene expression profiles, in order to retrieve biologically relevant information from the microarray data.

Proteomics

Chapter 6 describes an approach for diagnosis of disease from mass spectrometry data. The aim was to obtain a discrimination rule that best discriminates between breast cancer cases and controls. This endeavour was part of a classification competition, in which ten different research groups applied their approach. In our approach, ranking was performed by applying RF. Top-ranked variables showed to be highly correlated and were grouped into new variables. These newly created variables were finally included in a model to predict breast cancer cases and controls.

Applying partial least squares (PLS) as a multivariate statistical technique is shown in chapter 7 to conjointly analyze the association of 83 plasma proteins with CHD mortality and with intermediate endpoints known to play a role in CHD, namely BMI, HDL-C and total-C. PLS was applied to select a set of proteins with prognostic value for CHD mortality and to select sets of proteins associated with the intermediate endpoints. Subsequently, the proteins that were selected for the different endpoints by PLS, together with the intermediate endpoints were included in principal components analysis (PCA) [21] to interpret the relationships between identified proteins, intermediate endpoints and CHD mortality. The statistical results obtained in this study were subsequently biologically interpreted.

General discussion

In the general discussion the results from this project are discussed in a broader context. First the possibilities and limitations in analyzing large nutrigenomic datasets in relation to complex diseases are discussed, including the systems biology approach. This is followed by a discussion on the benefits and limitations of genomic research for public health. In the last section a future perspective on nutrigenomic research will be discussed.

References

1. Strachan T, Read AP: Human Molecular Genetics, Third edn. New York: Garland Science; 2004.
2. Kaput J, Rodriguez RL: Nutritional Genomics: Discovering the Path to Personalized Nutrition. Hoboken: John Wiley &Sons, Inc.; 2006.
3. Bellman R: Adaptive Control Processes. Princeton NJ: Princeton University Press; 1961.
4. Thornton-Wells TA, Moore JH, Haines JL: Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet 2004, 20(12):640-647.
5. Vapnik V: The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1998.
6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005, 102(43):15545-15550.
7. Betensky RA, Louis DN, Cairncross JG: Influence of unrecognized molecular heterogeneity on randomized clinical trials. J Clin Oncol 2002, 20(10):2495-2499.
8. Afman L, Muller M: Nutrigenomics: from molecular nutrition to prevention of disease. J Am Diet Assoc 2006, 106(4):569-576.
9. Geladi P, Kowalski BR: Partial least-squares regression: a tutorial. Analytica Chimica Acta 1986, 185:1-17.

10. Boulesteix AL, Strimmer K: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 2007, 8(1):32-44.
11. Moore JH: Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 2004, 4(6):795-803.
12. Wilke RA, Reif DM, Moore JH: Combinatorial pharmacogenetics. Nat Rev Drug Discov 2005, 4(11):911-918.
13. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 2006, 241(2):252-261.
14. Jakulin A, Bratko I: Analyzing attribute dependencies. Lect Notes Artif Intell 2003, 2838:229-240.
15. Jakulin A, Bratko I, Smrke D, Demsar J, Zupan B: Attribute interactions in medical data analysis. Lect Notes Artif Intell 2003, 2780:229-238.
16. Moore JH, Williams SM: Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 2005, 27(6):637-646.
17. Li C, Li M: GWAsimulator: a rapid whole-genome simulation program. Bioinformatics 2008, 24(1):140-142.
18. Dudek SM, Motsinger AA, Velez DR, Williams SM, Ritchie MD: Data simulation software for whole-genome association and other studies in human genetics. Pac Symp Biocomput 2006:499-510.
19. Albers CJ, Jansen RC, Kok J, Kuipers OP, van Hijum SA: SIMAGE: simulation of DNA-microarray gene expression data. BMC Bioinformatics 2006, 7:205.
20. Coombes KR, Koomen JM, Baggerly KA, Morris JS, Kobayashi R: Understanding the characteristics of mass spectrometry data through the use of simulation. Cancer Inform 2005, 1:41-52.
21. Jackson JE: A user's guide to principal components, First edn. New York Wiley – Interscience; 1991.

Chapter 2

The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases

A Geert Heidema[1,3,4]
Jolanda MA Boer[1]
Nico Nagelkerke[2]
Edwin CM Mariman[3]
Daphne L van der A[1]
Edith JM Feskens[1,4]

[1] Centre for Nutrition and Health, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands;
[2] Department of Community Medicine, United Arab Emirates University, Ain, UAE;
[3] Department of Human Biology, Maastricht University, Maastricht, The Netherlands;
[4] Division of Human Nutrition, Wageningen University and Research Centre, Wageningen, The Netherlands.

BMC Genetics 2006, 7:23.

Abstract

Genetic epidemiologists have taken the challenge to identify genetic polymorphisms involved in the development of diseases. Many have collected data on large numbers of genetic markers but are not familiar with available methods to assess their association with complex diseases. Statistical methods have been developed for analyzing the relation between large numbers of genetic and environmental predictors to disease or disease-related variables in genetic association studies.

In this commentary we discuss logistic regression analysis, neural networks, including the parameter decreasing method (PDM) and genetic programming optimized neural networks (GPNN) and several non-parametric methods, which include the set association approach (SAA), combinatorial partitioning method (CPM), restricted partitioning method (RPM), multifactor dimensionality reduction (MDR) method and the random forests (RF) approach. The relative strengths and weaknesses of these methods are highlighted.

Logistic regression and neural networks can handle only a limited number of predictor variables, depending on the number of observations in the dataset. Therefore, they are less useful than the non-parametric methods to approach association studies with large numbers of predictor variables. GPNN on the other hand may be a useful approach to select and model important predictors, but its performance to select the important effects in the presence of large numbers of predictors needs to be examined. Both SAA and RF are able to handle a large number of predictors and are useful in reducing these predictors to a subset of predictors with an important contribution to disease. The combinatorial methods give more insight in combination patterns for sets of genetic and/or environmental predictor variables that may be related to the outcome variable. As the non-parametric methods have different strengths and weaknesses we conclude that to approach genetic association studies using the case-control design, the application of a combination of several methods, including SAA, MDR and RF, will likely be a useful strategy to find the important genes and interaction patterns involved in complex diseases.

Background

The field of genetic epidemiology aims to identify genetic polymorphisms involved in the development of diseases. Single-locus methods measure the effect of one locus irrespective of other loci and are useful to study genetic diseases caused by a single gene, or even loci within single genes. To study complex diseases such as cardiovascular disorders or diabetes single-locus methods may not be appropriate, as it is possible that loci contribute to a certain complex disease only by their interaction with other genes (epistasis), while main effects of the individual loci may be small or absent [1]. Single-locus methods can not detect complex patterns [2], thus underestimate the genetic contribution to disease in the presence of interactions between loci. Therefore, approaches have been developed that take into account that complex diseases can be caused by an intricate pattern of genetic variants. These approaches are referred to as multi-locus methods and are specifically designed to find multiple disease loci, possibly on different chromosomes [3]. Diseases with a polygenic background can be studied by multi-locus methods, but also multi-factorial diseases by incorporating environmental predictors into the model.

Studying the effect of multiple genetic and/or environmental predictors and their interactions is fraught with statistical problems. One of these problems involves multiple testing. For each tested locus the probability to make a type I error is present, which is the probability to accept the hypothesis that the locus has an effect while in reality it does not. By testing multiple markers independently the type I error probability of finding a false positive result is increased. Two correction procedures for multiple testing are Bonferroni procedure and the false discovery rate [4]. Adjusting for multiple testing leads to a decrease of power (the probability to detect an effect when the effect is present) which makes it less likely to find weak genetic effects. Several multi-locus methods, discussed later in this commentary, have been developed to solve the multiple testing problem. These methods have greater power to detect susceptibility loci than single-marker tests.

The problem of modest sample sizes to test interactions for a large group of predictors (high-dimensional data) is referred to as the 'curse of dimensionality' problem [5]. The number of observations becomes too small relative to the number of predictors tested as few or no observations for combinations of predictors will occur. Traditional parametric approaches suffer from the dimensionality problem as it results in inaccurate parameter estimates for interaction effects [6]. Multi-locus methods are needed to select from the large amount of genetic and environmental predictors a small group of predictors and/or interactions between predictors that have a significant effect on the disease outcome. Subsequently, parameters for the selected predictors can be estimated by logistic regression analysis.

A third problem in the analysis of the effect of multiple genetic and environmental predictors on disease is the presence of correlated predictors in the dataset. An example is the presence of SNPs that are in linkage disequilibrium (LD) among the set of SNPs tested for association with disease. The power of a method to detect important predictors can be decreased when correlated predictors are tested. Some of the multi-locus methods discussed in this commentary are able to handle correlated predictors. Very high correlations between predictors, which is referred to as multicollinearity, is always a problem for methods: highly correlated predictors have an equal chance to be selected and one predictor may

falsely be selected instead of the highly correlated predictor that is truly associated with disease. Multicollinearity can be coped with statistically by combining data from multiple predictors into a single variable [7], for example combining SNPs that are in high LD into haplotypes.

Another difficult problem is the presence of heterogeneity [8]. Genetic heterogeneity is present if different genetic loci are independently associated with the same disease. The genes in which these loci are present can be part of different etiological pathways leading to the same disease or be part of the same pathway. Irrespective of the biological mechanism that gives rise to genetic heterogeneity, the association of these loci with the disease will be reduced if the total sample is used for measuring the association. A method that is not robust in the presence of genetic heterogeneity will likely suffer from a decrease in power to detect genetic effects. If genetic heterogeneity is not handled it can be accounted for by employing cluster analysis of genetic markers to identify groups of individuals with similar genetic profiles [8]. If clusters are present, association analyses of markers with the outcome variable should be accommodated for cluster effects [9]. Another form of heterogeneity that can affect the power to detect markers associated with disease is the presence of phenocopies. Phenocopies are individuals affected by the disease while they have a low-risk genotype profile. These individuals have developed the disease due to certain environmental factors. As in the presence of genetic heterogeneity, phenocopies will decrease the association between genetic markers and the disease if the association is studied using the total sample. Cluster analysis of environmental factors in the population can be used to define subgroups and cluster effects should be taken into account in the association analyses.

Many genetic epidemiologists have collected data on large numbers of genetic markers but are not familiar with the available methods to assess their association with complex diseases. In this article we review the strengths and weaknesses of methods for analyzing the genetic and/or environmental effects on disease or disease-related variables. These methods are presented in figure 1. Logistic regression and neural networks are discussed so as to compare non-parametric methods with these more 'traditional' statistical methods. The non-parametric methods have been selected as several genetic association studies have been conducted using these methods to analyze their data. This field is rapidly in progress and more methods are becoming available. This commentary does not pretend to cover all available multi-locus methods, nor to provide their statistical background, but aims to function as a starting and reference point for researchers in the field of genetic epidemiology who want to become more acquainted with multi-locus methods.

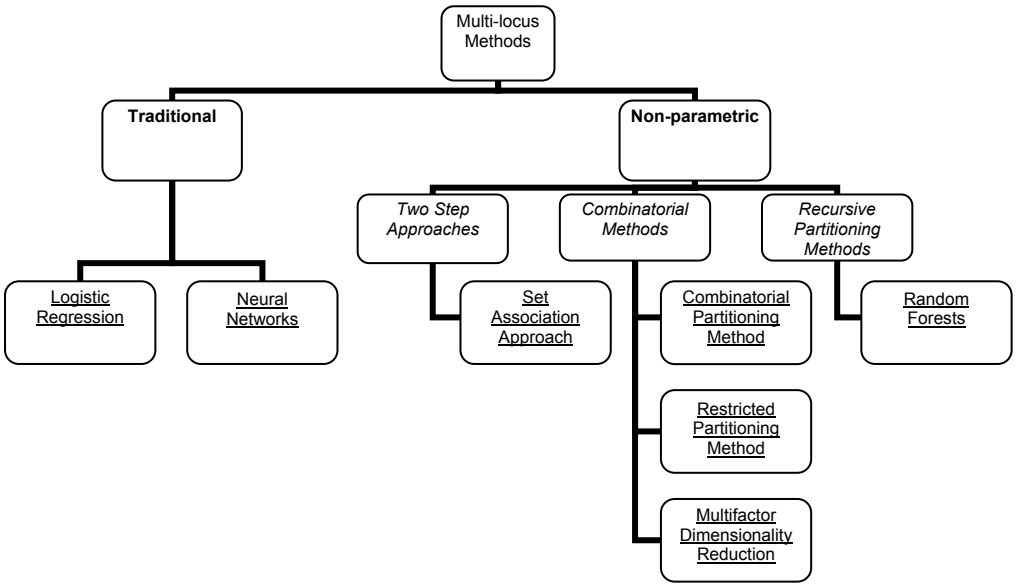


Figure 1: Diagram containing the different methods described in this commentary.

Method of evaluation

To provide an overview of the strengths and weaknesses of each method the ability of the methods to model the effects of multiple genetic and/or environmental predictors on disease outcome is evaluated for the following features:
- are the methods able to handle large numbers of predictors relative to the number of observations (dimensionality problem);
- number of predictors that can be analyzed in modest sample sizes;
- power to detect genetic effects;
- how do the methods handle interactions;
- do the methods maintain power if correlated predictors are present in the dataset;
- performance of the method if genetic heterogeneity is present;
- software availability and whether available software is open-source.

For each method a description is given followed by the discussion of the performance of the method for the different features. First, logistic regression is discussed, followed by neural networks, the set association approach, the combinatorial methods and the random forests approach. After the discussion of the different methods, strengths and weaknesses of these methods are compared and a strategy to analyze the effect of multiple genetic and environmental predictors on disease is proposed. As many of these methods use permutation testing to determine the statistical significance of predictors a short explanation of this test will be given here.

For testing the statistical significance of the association between selected predictors and the outcome variable permutation tests are used to obtain the distribution of the test-statistic under the null hypothesis of no association. Permutation tests generate many samples for

which the association between the predictors and the outcome variable has been disrupted by randomly distributing the values of the predictors or outcome variable over the observations. For each permuted sample the method is applied to calculate the test-statistic and together these test-statistics form the distribution of the test-statistic under the null hypothesis. The proportion of permutation samples with a value exceeding the value of the test-statistic of the observed data gives the significance level for the observed test-statistic [10].

To evaluate the ability of a model to classify and predict a certain outcome variable, multi-fold cross-validation is often used. This procedure will be explained because different methods use multi-fold cross-validation to obtain the classification and prediction error of models relating predictors to a certain outcome variable. In multi-fold cross-validation the data are randomly divided into groups of approximately the same size. The parameters of the model are estimated by all groups except for one, this remaining group is used for obtaining the prediction error (or prediction accuracy) of the model. As an example, a ten-fold cross-validation divides the data into ten groups of equal size. Nine groups are used to build the model. For quantitative traits, a fraction of the prediction error of this model is computed by the remaining group. By turns the ten groups are used to compute a fraction of the prediction error and the sum of the ten fractions forms the prediction error. For categorical outcome variables (e.g. disease status) the prediction error is calculated for each of the ten groups. To reduce arbitrariness in the division of the data into the different groups when estimating the expected (average) prediction error, the multi-fold cross-validation is repeated several times. Each time the data is randomly divided into the same number of groups. For quantitative traits the sum of the prediction errors obtained by the different cross-validations divided by the number of repeats gives an average prediction error. For dichotomic traits the average prediction error is the sum of prediction errors over all groups divided by the number of groups. The average prediction error is an unbiased estimate of the prediction error of the model.

Traditional methods

Logistic regression

A parametric statistical method often applied in genetic epidemiology is logistic regression. It is used to analyze the effect of genetic and environmental predictors on a dichotomic outcome, for example disease status. Predictors are linked to the outcome variable by the logit function. While many methods can be used to test for an association between predictors and disease in case-control studies, in such case-control studies logistic regression is the only appropriate method to consistently estimate the strength of association between a predictor and disease [11]. The conditional logistic regression (CLR) method is appropriate if stratification is present in the data, for example in a study design with matched cases and controls. CLR adjusts for the matching of the cases and controls by stratifying the matched case-control pairs.

Features of the logistic regression method

One of the disadvantages of the logistic regression method is that it performs poorly in the presence of the dimensionality problem; it may lead to false positive results [12] and a low power to detect interactions [6]. This may be overcome by stepwise regression analysis, which reduces the large number of predictors to a smaller number of predictors that are significantly related to disease. With forward selection, significant main effects and interactions between these main effects are included in the model. With backward selection, non-significant effects are excluded from the full model containing all parameters. There are drawbacks to the use of these standard selection procedures. With forward selection, interactions can only be tested for the main effects included in the model. Backward selection has the disadvantage that it cannot properly work in the presence of too many variables relative to the number of cases. Even if it does work, inclusion of too many parameters reduces the power of the model. Applying the least absolute shrinkage and selection operator (LASSO) [13] for selection of predictors in logistic regression may be more useful than standard selection procedures. This procedure shrinks the coefficients of predictors that are not important to zero, thereby selecting a subset from a larger number of predictors. It appears to have a better performance than standard backwards selection, but one disadvantage of the LASSO may be that it does not reduce the number of predictors substantially [14]. Therefore, for selection of important predictors it will also be useful to apply other selection methods before using logistic regression analysis to estimate the strength of association between selected predictors and disease.

Correlation between predictors may be a problem for logistic regression as different model building strategies may lead to different results [7]. Also, logistic regression does not handle genetic heterogeneity well as it models the relation between predictors and risk of disease for all individuals in the sample [15] and therefore it does not account for the presence of subgroups with different relationships between disease and genetic make-up. If different subsets of genes work in different subsets of the sample then logistic regression will probably not detect the different genetic causes of disease [16]. To perform logistic regression analysis many standard software packages (e.g. SAS, SPSS) are available.

Neural networks

Artificial neural networks are used to recognize patterns in the observed data and can be applied to determine genetic and environmental predictors related to disease. In genetic epidemiology, neural networks can be used to select SNPs that may contribute to disease. In this section we will describe in the first part the structure of a network that is commonly used (the feed forward network) and how neural networks usually are applied to obtain the best structure. In the second part we describe the parameter decreasing method [17], which can be used to select a subset of important predictors among a larger set of predictors. The genetic programming optimized neural network (GPNN) [18] is a strategy that will be described separately in the third part as it optimizes the structure of the network in a different way and different steps are involved to select the best model.

Structure of the feed forward network

A type of network commonly used consists of an input layer, one or more hidden layers and an output layer. Each layer is built up of nodes whereby one layer of nodes is connected to the next layer and weights are assigned to the connections. For example, with 10 input nodes, 6 hidden layer nodes and 1 output node the number of connections, and thus weights, equals 10*6 + 6*1 = 66. This type of network has a feed-forward structure: the flow of information is from the input layer, via the nodes of the hidden layer(s) to the node(s) of the output layer. The values of the predictors are the input values for the neural network.

The combined input values are processed by each of the nodes of the hidden layer by a transfer function. For dichotomic outcome variables the transfer function is for example the logistic function. A network containing one hidden layer node with a logistic transfer function is equivalent to logistic regression analysis [19] and networks containing more hidden nodes with logistic transfer functions are generalizations of logistic regression to more complex non-linear relationships between predictors and disease [20]. These non-linear relationships do not need to be defined. More layers and nodes increase the complexity of the model which enables the network to model complex interactions between the predictor variables. Networks fall in between parametric and non-parametric approaches as they provide large but not unlimited numbers of parameters to analysis methods [20].

The output of each node is determined by the outcome of the transfer function and is processed by each node of the next hidden layer (if present). The output of the last hidden layer is processed by the output node. The network associates the input values of the predictors with the output values given by the network. The amount of error between the output values of the model and the observed values is measured by an error function, for example a sum squared error.

Training the network, i.e. essentially estimating all the (hidden) parameters in the transfer function, is the process of adjusting the weights of the connections whereby weights are increased if they improve the output values and decreased if they result in more error. The procedure to optimize the weights is referred to as the back propagation algorithm [19]. The aim of the training is to obtain the model containing weight values that minimize the classification error of the network. Multi-fold cross-validation is used to divide the data into a training set and an evaluation set. The network model is constructed using the training set, the evaluation set is used to obtain the prediction error of the model. The error between the predicted values and observed values of the evaluation set gives the prediction error of the network. Each group created by multi-fold cross-validation is used to obtain the prediction error and the average prediction error is given by the sum of the prediction errors divided by the number of groups. The best model is the model with the lowest classification and prediction error. After the model has been obtained, predictors associated with the disease can be selected.

Parameter decreasing method

To select important SNPs from the total group of SNPs that were used to construct the network model, a parameter decreasing method (PDM) can be used [17]. The procedure of

PDM starts by deleting one SNP from the total number of SNPs and constructs a model containing the remaining SNPs. In turn each SNP is deleted from the total number of SNPs and with the remaining SNPs a model is constructed. From the constructed models the model with the lowest number of misclassified subjects in both the training and evaluation set is selected. This process is repeated until one SNP remains. For each selected model a measure of prediction accuracy is calculated by the sum of true predicted cases and controls divided by the total number of the evaluation sample. The prediction accuracy is calculated for each evaluation set created by multi-fold cross-validation and the sum of the prediction accuracies divided by the number of evaluation sets gives the average prediction accuracy.

The PDM has been applied to select from 25 SNPs a subset of susceptible SNPs of childhood allergic asthma [17]. The average prediction accuracy started to decrease after SNPs were excluded from the model containing 10 SNPs. To minimize the effect of randomized initial weight values, five PDM trials were performed and the importance of SNPs that remained in the last 10 SNPs of each trial was determined. For each trial the 10 SNPs were ordered from 1 to 10, based on the significance level of each SNP with the disease. The sum over the five trials for the SNPs that remained in the different trials was computed (sums can range from 1 to 50) and it is assumed that SNPs with higher scores are more important. The selected SNPs were used to construct models in order of importance of SNPs and for each model the prediction accuracy was calculated. Models with 10 of the most important SNPs or more had high prediction accuracy. The model containing the 10 most important SNPs had the same prediction accuracy as the model containing all 25 SNPs.

A permutation test can be applied to determine whether at least one of the selected SNPs is associated with the disease by randomly permuting the values of the selected SNPs [21]. To investigate important interactions, Tomita et al. [17] computed for 2-SNP and 3-SNP combinations the p-values by χ^2-test and selected SNP combinations with a p-value lower than 0.05. Combinations obtained in this manner likely contain false positive results because correction for multiple testing has not been applied. Therefore they used another measure of evaluation which they refer to as the effective combination value (ECV). If SNPs in a combination are independent, then the product of their separate p-values is equal to the p-value of the combination. ECV is the ratio of a SNP combination p-value divided by the product of SNP p-values and ECV < 1 suggests that interaction is present. SNP combinations that meet criteria for both χ^2 p-values and ECV values are selected.

Genetic programming optimized neural networks

A different strategy which can be used to select predictors associated with disease is referred to as genetic programming optimized neural networks [18]. Ritchie et al. [18] developed this strategy to optimize the neural network structure in order to improve selection of disease associated predictors. The back propagation algorithm described in the first part of the neural network section optimizes the weights. GPNN on the other hand not only optimizes the weights, but also a set of inputs that is selected from a larger set of predictors, the number of hidden layers and the number of nodes within the hidden layer(s). Cross-validation is also applied in GPNN to obtain for each partition of the data the best model and the prediction error for this selected model.

The genetic programming procedure starts with random selected models and evolves during the process to the model with the best structure. The steps taken by GPNN to obtain the best model are described here, more detailed information can be found in [18, 22]. First, a sample of all possible different GPNN models is randomly generated, using for each model a random subset of predictors from the total number of predictors. These initial GPNN models may differ in size. For each of the generated models is determined how well it fits the data, for example by its classification error. From these models a new generation of models is formed, which is equal to the number of models that were generated at the start of the process. This new generation of models is formed by directly copying a predefined proportion of the best models (those with the lowest classification error if classification error is used as fitness function) as well as by exchanging different parts between the models for another subset of best models. Thus, compared to the previous generation the new generation consists of similar models (the best proportion of models of the previous generation) and new models that are the result of recombining models of the previous generation (which is another subset of best models than the models that were copied). The size of the recombined models is allowed to change. The new generation of models replaces the previous generation and the process is repeated, bringing forth a next generation of models. This process continues until GPNN reaches a certain criterion (for example a classification error of zero or the maximum number of generations specified by the researcher). The model in the last generation that has the best fit (e.g. lowest classification error) is denoted as the best GPNN model and the prediction error for this model is determined by the remaining part of the data. For each partition created by cross-validation a best GPNN model with the corresponding prediction error is obtained. For example, 10-fold cross-validation will result in 10 best GPNN models.

To determine the importance of predictors or predictor combinations, a cross-validation consistency measure can be used, which is the number of times a predictor or predictor combination is selected in a best model across all validation sets, divided by the number of validation sets. The predictor or predictor combination which has the highest cross-validation consistency is denoted as the final selected model. An example of GPNN application to case-control data is the study of Motsinger et al. [22] on Parkinson's disease.

Features of neural networks

The advantage of neural networks over logistic regression is the possibility to flexibly model complex relationships between the predictor variables and the disease status. Tomita et al. [17] compared the prediction accuracy of constructed models of neural networks with logistic regression analysis for the models containing 25 and 10 SNPs. Constructed models by neural networks had high prediction accuracy while the accuracy was low for logistic regression analysis. A disadvantage of the PDM is that a cut-off value for the prediction accuracy to select SNPs as susceptible is not given.

In general, as the network can handle a limited number of predictor variables depending on the number of observations in the dataset, faced with testing very large numbers of genetic markers the network is subject to the dimensionality problem [3]. GPNN however is not subject to the dimensionality problem because it uses only a random selection of

predictors to build the initial GPNN models and selects the most important predictors during the process.

Studies investigating the power for neural networks using PDM have not been found in the literature, thus information about the power of the PDM to detect important effects is not available at the present time. For GPNN, the power to detect important SNPs in the presence of unrelated SNPs is higher compared to the commonly used feed forward NN using a back propagation algorithm [18]. Using simulated data, Motsinger et al. [22] showed that the power of GPNN to detect gene-gene interactions in two and three locus interaction models is high. The number of unrelated SNPs included however was not large and further information on the power of GPNN to detect genetic effects among a large set of unrelated SNPs is needed.

If important interactions between SNPs are present, PDM will likely be able to detect the SNPs involved in the interaction, because deleting a SNP would have an effect on the prediction accuracy. Important interactions between SNPs will therefore lead to selection of these SNPs. Also, most of the 2-SNP and 3-SNP combinations identified by Tomita et al. [17] were combinations of SNPs that had been selected by the PDM procedure, followed by combinations of selected and unselected SNPs and the least number of combinations was found for unselected SNPs. This suggests that neural networks are able to select SNP combinations accurately [17].

For detection of the genetic polymorphisms involved in disease, correlated markers are a problem for neural networks using the PDM. If one marker is associated with disease, but is correlated with another marker, deleting the marker associated with disease will result in a smaller decrease in the value of the prediction accuracy compared to a situation of uncorrelated markers. Therefore, the power to detect the association of the risk marker with the disease will be reduced if this marker is correlated with one or more other markers. The power of GPNN to detect important predictors will not be reduced when correlation between predictors is present. GPNN models containing important predictors are more informative and will have lower classification errors than models containing predictors correlated with the important predictors. Important predictors will therefore be selected during the process.

Neural networks can determine substructures within a dataset which enables them to handle genetic heterogeneity [20]. Software to perform neural network analysis of case-control data using PDM is freely available [23], the software is however not open-source. At the moment, software for GPNN is not available.

Non-parametric methods

Two step approaches

Several genetic association studies have employed the two step approach, which consists of the following two steps:

- Step 1: determine a small number of potentially important markers;
- Step 2: model interactions between important markers and/or environmental predictors.

In the first step a non-parametric approach is applied to reduce many markers to a small number of important markers. For the second step environmental predictors can be

introduced to the model and logistic regression or neural networks can be used to test gene-gene and/or gene-environment interactions. In the two step approach coupled-logistic regression can be applied to analyze interactions between the selected markers obtained in the first step [24, 25]. The coupled-logistic regression procedure first uses one forward selection step to model the two-way and higher-order interactions between the selected markers and environmental predictors if included. Then backward selection is employed to eliminate non-significant interactions.

Set association approach

A non-parametric approach for selecting a set of important markers as a first step is the set association approach (SAA). SAA is described in this section; more detailed information can be found in [26]. Instead of categorical predictors such as marker genotypes, SAA can also be used to analyze the effect of quantitative predictor variables [27].

SAA starts by calculating a test-statistic for each marker separately, which is a product of two test-statistics. The first statistic measures the association of a marker with disease outcome. As measure of association χ^2 can be calculated from the contingency table of alleles (or genotypes) with disease status, but other statistics can be used as well. The deviation of a marker from the null-hypothesis of Hardy-Weinberg (HW) equilibrium is used as the second test-statistic, which is chi-square distributed. χ^2 values for deviations from HW equilibrium are calculated in the case group and larger deviations indicate an association between the marker and the disease. Very large χ^2 values for HW disequilibrium in the control group can indicate genotyping errors. To correct for the quality of genotyping, markers showing large χ^2 values in controls (e.g. χ^2 values exceeding the χ^2 value corresponding to the 99-th percentile) are deleted or set to zero [26]. Thus, for the calculation of the test-statistic for each marker, information is used from allelic association, deviation from HW equilibrium and genotyping errors.

Subsequently, the markers are ordered based on their value for the test-statistic. SAA starts with the selection of the marker with the largest test-statistic and calculates sum-statistics by adding each time the most important marker from the group of unselected markers. Increasing sums of markers are formed and the number of markers in the sums ranges from 1 to a predefined maximum number of M markers, for example 20. The significance level of each sum of markers is tested using a permutation test. SAA uses, and holds fixed, the observed genotypes, but randomly permutes the variable that indicates the disease status. Many permuted samples are formed and for each sample sum-statistics are calculated. The p-value for a certain sum of markers represents the proportion of permuted samples exceeding the value of the sum of markers of the observed sample. Instead of testing many markers, M sums (e.g. 20) are tested. Increased number of markers in the sum with an association with disease will lower the significance level of the sum. At a certain point the significance level will no longer decrease but increase as markers not contributing to disease are added to the sum. Therefore, from the M sums tested the set of markers with the lowest significance level is selected as the best set of markers. This p-value is defined as test-statistic and is evaluated by a second permutation test testing the null-hypothesis of no association of the selected markers with the disease-outcome. The second round of

permutation results in an overall p-value reducing the testing of M sums to one sum. The multiple testing problem that arises due to testing many markers has been overcome at this stage of SAA.

Applications of SAA have been reported for case-control studies on heart disease [27] and Alzheimer's disease [28].

Features of the set association approach

SAA manages the dimensionality problem by reducing the number of markers to a smaller number of important markers. This method also provides an overall significance level for the selected markers. In general, the main advantage of two step approaches is that large numbers of markers can be evaluated for their importance in contributing to disease. Compared to the Bonferonni and the False Discovery Rate procedures that correct for multiple testing, SAA has more power to identify genes involved in disease; sum-statistics are compounds of marker main effects which have a better performance than approaches that test each marker independently [29]. Furthermore, the power of SAA is enhanced by using information from allelic association, deviation from HW equilibrium and genotyping errors.

The main disadvantage of SAA is that genetic interactions are only tested for the markers that are selected in the sum. Important interactions with weak main effects will be missed.

To handle correlations between markers, Wille et al. [29] proposed a method to adjust the test-statistic of a marker for the correlation with the markers that are already present in the sum. Using unadjusted test-statistics, markers could be included in the sum while these markers are correlated with markers already contained in the sum. If the correlation between markers concerns non-susceptibility loci it will result in an overrepresentation of non-susceptibility loci in the sums, reducing the power of the approach. By using adjusted marker statistics the power of the test is re-established.

Genetic heterogeneity will affect the performance of SAA to identify important markers as this approach tests the association of markers with disease for the whole sample. Individuals affected due to different loci decreases the association between each of the loci with the disease and will result in a reduction of power of SAA to detect these loci. SAA is implemented in the program Sumstat [26] which is freely available and open-source [30]. At the moment, the adjustment of marker statistics for their correlation with markers already included in the sum has not been implemented in the software.

Combinatorial methods

Combinatorial methods search over all possible factor combinations to find combinations with an effect on an outcome variable. The combinatorial methods that will be discussed are the combinatorial partitioning method (CPM), the restricted partitioning method (RPM) and the multifactor dimensionality reduction method (MDR). Respectively, CPM and RPM have been described more extensively by Nelson et al. [31] and Culverhouse et al. [32]. Several recent reviews are available for MDR [33, 34]. CPM and RPM aim to identify factor combinations that explain best the variance of a quantitative phenotype. MDR classifies factor combinations as having a low risk or high risk on disease based on the

presence of these combinations in cases versus controls. Both CPM and MDR use multi-fold cross-validation to select the factor combinations that have the best prediction of the outcome variable and to compute the average proportion of variability explained (CPM) or average prediction accuracy (MDR), which is used to evaluate the validity of the obtained factor combinations. It is important to evaluate the validity of the model to verify whether the combinations do not present false positive results but are truly associated with the disease [35].

Combinatorial partitioning method

CPM can be used to study the effect of factor combinations on a quantitative phenotype. This phenotype can be a variable underlying the disease of interest. An example is to study factor combinations involved in the phenotype blood pressure, which underlies cardiovascular disease. To test whether a locus has an effect on a quantitative phenotype, analysis of variance (ANOVA) could be used. It performs an overall test of the differences between the mean phenotypic values of genotypes. However, with many genotypes a posteriori testing the significance of the differences between genotype means leads to the problem of multiple testing. CPM has the advantage that it determines the loci combinations with an effect on a quantitative phenotype and at the same time defines groups of genotypes with similar phenotypic means [31]. In the CPM a group of genotypes with similar phenotypic means is referred to as a genotypic partition. Combinations of two or more partitions make up a set of genotypic partitions. CPM selects sets of genotypic partitions (consisting of multi-locus genotypes) that predict variation of the quantitative trait [31]. The CPM consists of three steps:
- Select loci combinations from all loci studied. For these loci combinations, combine genotypes with similar phenotypes into partitions. Select from the total group of partitions each combination of genotypic partitions (thus each set) that predicts a certain level of variance;
- Validate each selected set by multi-fold cross-validation;
- Select the most predictive sets and make inferences about the combinations of loci and the genotype-phenotype relationships.
In the first step the combinatorial partitioning method selects all possible subsets of loci from the total group of loci that is studied. For example, if 10 loci are studied and all 2-loci combinations are considered, the number of subsets of loci examined is equal to $\binom{10}{2} = 45$ pair wise combinations. For each subset of loci all genotypic partitions are examined. For two SNPs at autosomal loci the number of genotype combinations equals nine and the number of genotypic partitions investigated ranges from two till nine. A set can for example consist of two genotypic partitions, one partition containing the multi-locus genotypes AAbb, AaBb and aabb and the other partition containing the remaining genotypes. CPM evaluates all possible sets of genotype partitions and selects sets based on two criteria. The first criterion is the proportion of phenotypic variability explained by a set. For each selected set the variability between partitions should be much higher than within the partitions because these sets will explain the largest proportion of variance of the quantitative phenotype. The other criterion used is the number of individuals in a set. Only

a few individuals will be present for genotypes with low frequency alleles and consequently for partitions in which these genotypes are present. Small numbers for genotypic partitions leads to unreliable estimates of the partition means and partition variance. When the number of individuals is set too low, spurious effects may be found by chance. On the other hand, genotypic partitions that do have an effect could be discarded from further analyses when the number of individuals is set too high [31].

In the second step each selected set is validated by the multi-fold cross-validation method. For validation of the selected sets all the groups generated by the cross-validation method, except for one, are used to estimate the means of the genotypic partitions of a set. The remaining group is used to compute the within partition sum of squares, which is the predicted error for this group only. The sum of the fractions of the different groups gives the total predicted error of a set. If the multi-fold cross-validation is repeated several times, an average predicted error can be calculated. From this averaged predicted error the proportion of variability explained by a set is computed, which is a measure of the predictive ability of the phenotype by a set of genotypic partitions. Sets with smaller proportions of within sum of squares explain more variability of the quantitative phenotype and thus have a higher predictive ability.

Based on the results of the cross-validation, the most predictive sets are selected in the third step. It is useful to select more than one predictive set of genotypic partitions, because by comparing the different sets more insight in the relation between combinations of loci present in these sets and the quantitative trait can likely be gained. To obtain the statistical significance of the most predictive set selected a permutation test can be performed. Phenotypic outcomes are randomly assigned to the genotypes and for each permutation sample the CPM is performed. The null-hypothesis tested is that the most predictive set is not significantly associated with the quantitative trait. The proportion of sets exceeding the observed value of proportion of variability explained by the most predictive set results in a p-value for the most predictive set.

CPM has been applied in studies of plasma triglyceride levels [31], plasma PAI-1 levels [36] and the relationship between plasma t-PA and PAI-1 levels [37].

Restricted partitioning method

To overcome the computationally intensive search technique used by CPM, Culverhouse et al. [32] developed the restricted partitioning method. Where CPM searches over all possible combinations, RPM restricts its search in order to avoid evaluation of genotype partitions that will not explain much of the variation. The reasoning is that a group consisting of genotypes for which the difference between their mean values is large (thus having a large within group variance), will not explain much of the total variance of the quantitative trait and can therefore be discarded for evaluation.

The search procedure that is used by RPM to select genotypic partitions consists of the following steps:
- Using a multiple comparison test, examine whether significant differences between mean values of genotype groups are present (at the start of the analysis each group consists of one multi-locus genotype);
- from all the non-significant pairs of genotype groups, combine the pair with the smallest difference between their mean values into a new group, thereby reducing the number of genotype groups to be evaluated with one;
- the procedure is reiterated until all differences between pairs of genotype groups are significantly different.
If all the genotypes have significantly different means in the first step the procedure ends at this step. Otherwise, the number of genotype groups in the final partitioning is less than the number of genotypes present at the start of the analysis. To measure the importance of the final model R^2 is determined, which is the proportion of the trait variation explained by the genotype groups. The significance of the model is estimated by permutation testing, generating a null distribution of R^2. Bonferroni correction is applied for the number of factor combinations that have been tested. Factor combinations are selected if the explained variance R^2 by the combination is found to be significant. Analysis with RPM has been performed for irinotecan metabolism [32].

Multifactor dimensionality reduction method

The multifactor dimensionality reduction method analyzes genetic and/or environmental effects on a dichotomic outcome variable (e.g. disease status) rather than a quantitative trait. MDR has been inspired by the CPM, but the approach differs in many perspectives. From the total group of factors studied, MDR evaluates all possible N-factor combinations of genetic and/or discrete environmental factors. Each cell of the N-factor combination is assigned to either a low risk or high risk group. A certain threshold, defined as the ratio of cases to controls, determines the risk group to which a factor combination is assigned. For example, for all nine possible genotype combinations of each two loci combination the risk status is determined. If the threshold is set to one and the cell for a genotype combination contains more cases than controls, that genotype combination is determined as high risk. Thus, MDR assigns each combination (e.g. multi-locus genotype) within a N-factor combination to a high risk or low risk group, thereby constructing a new factor consisting of the two risk groups. The process of constructing a new factor as a function of two or more other factors is referred to as constructive induction and MDR can therefore be viewed as a constructive induction approach [38]. MDR evaluates the ability of this new factor to classify and predict disease status by multi-fold cross-validation.

Multi-fold cross-validation divides the observed data in equal subsets. One subset remains aside, the other subsets are used to build the model. The N-factor model with the lowest classification error is selected and for this model the remaining subset is used to obtain the prediction accuracy. By turns each subset is used to obtain the prediction accuracy for the best classifying model that has been build by the other subsets. The model with the highest prediction accuracy is selected as the best N-factor model.

Different numbers N of factors are evaluated. For each number of factors, multi-fold cross-validation is used to select the best classifying N-factor combination by measuring the prediction accuracy of the model. Cross-validation consistency (also discussed in the genetic programming optimized neural networks section) is another measure for selecting the best classifying N-factor combination: it is the number of times a N-factor combination is selected as the best model across all validation sets, divided by the number of validation sets. The N-factor model with the highest prediction accuracy and/or the highest cross-validity consistency is selected. If one best model is found with the highest prediction accuracy and another model with the highest cross-validation consistency, the most parsimonious model is chosen for describing the observed data. For example, if the best 2-factor combination model has the highest prediction accuracy and the best 3-factor combination model has the highest cross-validation consistency, the 2-factor combination model is selected. A permutation test is performed to obtain the statistical significance of the most predictive N-factor model. For each permuted dataset the best model is selected and the prediction accuracy or cross-validity consistency is determined. The p-value is obtained using the distribution of the prediction accuracy or cross-validity consistency under the null-hypothesis.

The MDR approach has been applied for example to case-control data of prostate cancer [39], type 2 diabetes [40], myocardial infarction [35], hypertension [41] and sporadic breast cancer [42].

Features of the combinatorial methods

The combinatorial methods discussed above select from all factor combinations those factor combinations that best explain the outcome variable thereby solving the dimensionality problem. Because both CPM and MDR are computationally intensive procedures the number of factors to be analyzed by these methods is moderate. Selection methods to preselect factors can be used as a first step [38] and such filter methods are part of the MDR software [43]. These methods can be applied before using MDR, enabling the user of the MDR software to analyze large numbers of factors. Although RPM has the advantage that it relieves the computational intensity of CPM and thereby has the potential to analyze many interacting loci, the multiple testing problem is still a challenge for this method.

One of the merits of the combinatorial methods is their high power to identify high-order interactions between loci while main effects are not present [32, 44]. The power of the MDR approach to detect gene-gene interactions in the absence of main effects was examined by Ritchie et al. [44]. Using simulated datasets, they studied 6 different models of interaction between two loci, including in the datasets noise due to 5 percent genotyping error, 5 percent missing data, 50 percent phenocopy and 50 percent genetic heterogeneity. Without noise factors, the power of the MDR method to detect the two-locus interaction for the 6 models was in between 80 and 100 percent. The drop in power due to genotyping errors, missing data or the combination of these noise factors was very small. Phenocopies had a large effect on the power for 4 models and genetic heterogeneity had the largest impact on the power for 5 of the 6 models. The power is reduced by phenocopies or genetic heterogeneity, because different combinations of factors causing the disease will decrease

the prediction accuracy and cross-validity consistency of a model [35]. The power to detect the interaction for each of the models was decreased to around 1 percent for the combination of phenocopies and genetic heterogeneity. If phenocopies are present, the power of the MDR approach can be increased if environmental factors causing the disease are included in the analysis. Environmental differences in the population can be assessed to define subgroups after which MDR can be applied to each group, or the environmental factors can be included in the MDR analysis. To account for genetic heterogeneity, cluster analysis of genetic markers can be employed (see background section). MDR analysis for the different clusters can be performed or the cluster status can be included as a covariate [44]. If the presence of genetic heterogeneity is not known beforehand the power for CPM, RPM and MDR is largely reduced.

As CPM, RPM and MDR select the model that has the best prediction of disease status, the model that contains the most information will be selected. Risk predictors contain more information than predictors correlated with the risk predictors and the power of these methods to detect risk SNPs will not be reduced when correlation between predictors is present. Software for CPM is not available, but a program for this method can easily be made by a competent statistical geneticist. Software implementing RPM is available from Culverhouse et al. [32]. Also, open-source software for RPM is currently under development [43]. MDR software, originally discussed by Hahn et al. [45], is freely available and is open-source [43]. A MDR Permutation Testing module to perform permutation testing is also freely available [43].

Recursive partitioning methods

Recursive partitioning methods partition the total dataset recursively into smaller and more homogeneous subsets to fit models for predicting the value of a continuous or categorical outcome from many predictor variables. These models are called tree-based models as the splits of the data into more and more homogeneous subsets can be pictured by a tree graph [15]. Regression and classification trees are respectively applied to continuous and categorical outcome variables. Here, the application of random forests (RF) of classification trees to case-control data is discussed.

A tree is made up of internal and terminal nodes, with the first internal node called the root node that contains the total sample. The root node is split into two nodes to improve the homogeneity of the case group and control group compared to the root node. This split is based on a cut-off point of the predictor variable that partitions the total sample best into the two groups of cases and controls, for example a split based on a certain SNP with one subset containing wild-type homozygous and heterozygous individuals (genotypes AA and Aa) and the other subset containing homozygous mutant individuals (genotype aa). Each of these two nodes is split again, whereby splits are based on the predictor variable that improves the homogeneity of the resulting subsets (this predictor may differ for each node). A node that is not further split into two nodes is called a terminal node. A recursive partitioning method that can be used for selection of important predictors contributing to disease is RF.

The random forests approach

In RF a group of tree-based models is used to select predictors with an important contribution to an outcome variable [16, 46]. For each model, every split is based on a random selected subset of all predictors studied. More important predictors will discriminate best between cases and controls and will therefore be closer to the root node and present in most of the trees. On the other hand, less important predictors will be less present in the different trees and closer to the terminal nodes [46]. RF has been described in more detail by Lunetta et al. [16] and Bureau et al. [46].

The prediction accuracy of the forest

For each tree in the forest, the total sample started with at the root node is generated by bootstrap sampling. With bootstrap sampling individuals are sampled from the observed population sample. The number of individuals in the bootstrap sample equals the number of individuals in the observed sample and because sampling is performed with replacement, some individuals can be present more than once in the bootstrap sample while other individuals are left out. The bootstrap sample is used to build the tree and the left-out individuals to obtain the prediction of the forest. The predictor values of a left-out individual determine which terminal node, or class, this individual is assigned to for a certain tree. The class to which most of the individuals of the bootstrap sample are assigned to is the predicted class of the tree for the left-out individual. The prediction for the forest is obtained by counting the predictions over the trees for which the individual was left out the bootstrap sample. The class with the most predictions is the prediction of the forest. In case-control data the prediction accuracy of the forest is given by the difference between the proportion of correct and incorrect classification of the left-out individuals. The prediction accuracy of the forest is used to obtain a measurement of importance of each predictor.

The importance of predictors

Predictors that best classify the population into cases and controls are assumed to be important predictors of disease-status. The importance of a predictor is given by an importance index I_m which denotes the importance of a predictor taken other predictors into account. The values of the predictor for which the index is computed need to be randomized for the left-out individuals to remove any association between the predictor and disease status. The importance index for predictor A is then the difference in prediction accuracy of disease status by the predictor vector and the same predictor vector with predictor A randomly permuted for the left-out individuals. Larger differences in prediction accuracy between the two predictor vectors indicate more important predictors. RF orders the predictors according to their importance. Computing the importance index can be extended to pairs of predictors whereby the predictor values of both predictors are permuted.

Application of RF to case-control data has been reported for asthma [46].

Features of the random forests approach

As RF selects the most important predictors among all predictors, the dimensionality problem is circumvented, but the approach does not provide a cut-off value of the importance index to determine which predictors should be retained for further analysis [16]. The advantage of RF is that it is able to test many predictors. Permuting the predictor values for the left-out individuals does not only remove the association between the permuted predictor and the outcome variable, but also the interaction effects of the permuted predictor with other predictors, if present. Thereby, the interactions of the predictor with other predictors are captured in the importance index. Lunetta et al. [16] tested the performance of RF compared to Fisher's Exact test in ranking risk SNPs using simulated data. Genetic heterogeneity was included in the disease models. If interaction between two risk markers is present, RF has a better performance to rank these risk markers than univariate ranking methods because the importance of each marker involved in the interaction will increase. More interactions and larger groups that interact increase the relative performance [16]. Therefore, markers with weak main effect but significant interaction with other markers can be detected by RF. The joint importance of subsets of predictors can be tested for all markers if the size of the subset is small, but testing the joint importance for larger subsets to capture higher-order interactions becomes computationally unfeasible. As Province et al. [15] point out, recursive partitioning methods are able to detect genetic heterogeneity. This assertion is confirmed by the study of Lunetta et al. [16]. Genetic heterogeneity is handled because different models are fitted to subsets of the data defined by early splits in the trees [15, 16]. Limited simulations suggest that correlated predictors are a problem for RF as it leads to a decrease of the predictor importance for each correlated risk SNP [16]. Software for RF is freely available and is open-source [47].

Conclusion

An overview of the strengths and weaknesses of the methods discussed is given in table 1. The dimensionality problem is not solved by the method of logistic regression. Applying a parameter decreasing method within neural networks to select important predictors is a useful approach if moderate numbers of SNPs are tested. However, neural networks can not handle the dimensionality problem either if the number of predictors tested becomes too large. Logistic regression and neural networks are therefore less useful to approach association studies with large numbers of predictor variables. These methods can be applied to model the effects of a group of selected predictors, including interaction terms and other potential risk factors. For example in the two step approach coupled logistic regression can be used after the markers have been selected in the first step. Genetic programming optimized neural network is able to select and model important predictors from a set of predictors, but the performance of GPNN to detect important SNPs in the presence of large numbers of unrelated SNPs needs to be investigated.

Table 1: Comparison of the different methods.

	Logistic regression	Neural networks		SAA	CPM	RPM	MDR	RF
		PDM	GPNN					
Type of outcome variable	dichotomous	categorical continuous	categorical continuous	dichotomous	continuous	continuous	dichotomous	categorical continuous
Dimensionality	no	no	yes	yes	yes	yes	yes	yes
Number of predictors	few	moderate	many	many	moderate	many*	moderate†	many
Power to detect important effects	low	no information	high	high	high	high	high	high
Detect interactions when main effects are absent	no	yes	yes	no	yes	yes	yes	yes‡
Correlated predictors	no	no	yes	not implemented**	yes	yes	yes	no
Genetic heterogeneity	no	yes	yes	no	no	no	no	yes
Software available	yes	yes	no	yes	no	at request	yes	yes
Open source		no		yes		and to be developed	yes	yes

For the problems of dimensionality, correlated predictors and genetic heterogeneity yes and no indicate respectively that a method is able or not able to handle the problem. For detection of interactions when main effects are absent yes and no indicate respectively that a method is able or not able to detect interactions while main effects of the loci involved in the interaction are small or absent.

* RPM is subject to the multiple testing problem.

† MDR can analyze a moderate number of factors, but filter methods that are part of the MDR software can be applied before using MDR, enabling the user of the MDR software to analyze large numbers of factors.

‡ Interactions contribute to the importance of predictors.

** Adjustment of the test-statistics for correlation between markers is not implemented in the software.

Both SAA and RF can handle a large number of predictors and are useful in reducing the large amount of predictors to those predictors with an important contribution to disease. Another argument for employment of RF is the possibility to detect the presence of genetic heterogeneity. The combinatorial methods are useful to give more insight in interaction patterns for sets of genetic and/or environmental predictor variables. CPM and RPM can be applied in the study of quantitative phenotypes underlying the disease of interest, MDR is useful for analyzing effects on disease status.

As each of the non-parametric methods has its strength and weaknesses, genetic association studies should be approached by several methods. For genetic association studies using the case-control design to analyze complex diseases, the application of SAA in combination with the MDR and RF will most likely be a useful strategy to find the important genes and interaction patterns involved, as each of these methods approach the analysis of multiple SNP data differently. Similarities and differences in the results generated by these methods will provide valuable information whether selected SNPs are likely to contribute to disease by their main effects or whether gene-gene interactions play a role. Thus the combination of these methods will give more insight in the etiology of complex diseases. These methods can also be used in a multi-step approach, discussed by Moore et al. [38], to detect and interpret interactions. In the first step of this approach a subset of important SNPs is selected from the total number of SNPs. SAA and/or RF could be applied as method for selection of important predictors. The next step is to apply a constructive induction approach to construct from this subset of SNPs a new factor consisting of high risk and low risk genotype combinations. MDR can be used at this second step as a constructive induction approach. The ability of this constructed factor to classify and predict disease status is evaluated in the third step, for example by multi-fold cross-validation which is also implemented in MDR. Besides detecting statistical interactions this multi-step approach provides the means to statistically interpret the detected interactions in the fourth step.At this last step visual tools can be used for model interpretion. This multi-step approach is flexible as at each step many different methods can be used [38].

More statistical methods to analyze multiple SNPs in relation to complex diseases are becoming available. What the features of other newly developed methods for analysis of multiple SNPs will be has to be studied and compared to the methods discussed in this commentary. Also, applications of the methods in genetic association studies will have to be performed in order to examine their practical value for the field of genetic epidemiology.

In this commentary the strengths and weaknesses of methods to approach the statistical challenge to detect gene-gene interactions associated with the disease or disease related outcome of interest have been discussed. However, these methods test interactions statistically, which is only a first step in the unravelling of the interacting underlying biological mechanisms. The biological interpretation of statistically detected gene-gene interactions is not straightforward and forms another challenge for genetic epidemiologists. Statistical interaction is detected on the population level by relating genotype information to interindividual differences in phenotype while biological interaction is the result of physical interaction of biomolecules which takes place at the individual level [2]. To address this challenge, Moore et al. propose the application of systems biology (a synthesis of multiple disciplines) to unicellular organisms, reasoning that understanding of the

relationship between statistical and biological interaction in these organisms will reveal some basic underlying principles and thereby will help to understand how statistical interaction is related to human complex diseases [2].

In conclusion, statistical methods have been developed that enable genetic epidemiologists to detect important genetic and/or environmental predictors associated with disease or disease related variables. These methods have different strengths and weaknesses. Applying a combination of these methods will provide insight in the main effects and interaction patterns involved in the etiology of complex diseases.

Acknowledgements

This project has been carried out within the framework of the Centre for Human Nutrigenomics.

References

1. Culverhouse R, Suarez BK, Lin J, Reich T: A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 2002, 70:416-471.
2. Moore JH, Williams SM: Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays 2005, 27:637-646.
3. Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003, 4:701-709.
4. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001, 125:279-284.
5. Bellman R: Adaptive Control Processes. Princeton NJ: Princeton UniversityPress; 1961.
6. Moore JH, Williams SM: New strategies for identifying gene-gene interactions in hypertension. Ann Med 2002, 34:88-95.
7. Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D: An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Prev Vet Med 1997, 29:221-239.
8. Thornton-Wells TA, Moore JH, Haines JL: Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet 2004, 20:640-647.
9. Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, Jacob HJ, Cohen D: The future of genetic case-control studies. Adv Genet 2001, 42:191-212.
10. Cox DR, Hinkley DV: Theoretical statistics London: Chapman and Hall; 1974.
11. Nagelkerke N, Smits J, Le Cessie S, Van Houwelingen H: Testing goodness-of-fit of the logistic regression model in case-control studies using sample reweighting. Statist Med 2005, 24:121-130.
12. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996, 49:1373-1379.
13. Tibshirani R: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society 1996, 58:267-288.
14. Li L, Huang J, Sun S, Shen J, Unverzagt FW, Gao S, Hendrie HH, Hall K, Hui SL: Selecting pre-screening items for early intervention trials of dementia – a case study. Statist Med 2004, 23:271-283.
15. Province MA, Shannon WD, Rao DC: Classification methods for confronting heterogeneity. Adv Genet 2001, 42:273-286.
16. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004, 5:32.

17. Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, Honda H: Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. Bioinformatics 2004, 5:120.
18. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH: Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics 2003, 4:28.
19. Bishop CM: Neural networks for pattern recognition Oxford: Clarendon Press; 1995.
20. Lucek PR, Ott J: Neural network analysis of complex traits. Genet Epidemiol 1997, 14:1101-1106.
21. North BV, Curtis D, Cassell PG, Hitman GA, Sham PC: Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. Ann Hum Genet 2003, 67:348-356.
22. Motsinger AA, Lee SL, Mellick G, Ritchie MD: GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics 2006, 7:39.
23. Software for Parameter Decreasing Method. [http://www.nubio.nagoya-u.ac.jp/proc/english/indexe.htm].
24. Ott J, Hoh J: Statistical multilocus methods for disequilibrium analysis in complex traits. Hum Mutat 2001, 17:285-288.
25. Zee RYL, Hoh J, Cheng S, Reynolds R, Grow MA, Silbergleit A, Walker K, Steiner L, Zangenberg G, Fernandez-Ortiz A, Macaya C, Pintor E, Fernandez-Cruz A, Ott J, Lindpaintner K: Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. Pharmacogenomics J 2002, 2:197-201.
26. Hoh J, Wille A, Ott J: Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 2001, 11:2115-2119.
27. Ott J, Hoh J: Set association analysis of SNP case-control and microarray data. J Comput Biol 2003, 10:569-74.
28. De Quervain DJF, Poirier R, Wollmer MA, Grimaldi LM, Tsolaki M, Streffer JR, Hock C, Nitsch RM, Mohajeri MH, Papassotiropoulos A: Glucocorticoid-related genetic susceptibility for Alzheimer's disease. Hum Mol Genet 2004, 13:47-52.
29. Wille A, Hoh J, Ott J: Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. Genet Epidemiol 2003, 25:350-359.
30. Software for set association approach [http://linkage.rockefeller.edu/ott/sumstat.html]
31. Nelson MR, Kardia SLR, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 2001, 11:458-470.
32. Culverhouse R, Klein T, Shannon W: Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 2004, 27:141-152.
33. Moore JH: Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 2004, 4:795-803.
34. Wilke RA, Reif DM, Moore JH: Combinatorial pharmacogenetics. Nat Rev Drug Discov 2005, 4:911-918.
35. Coffey CS, Hebert PR, Ritchie MD, Krumholz HM, Gaziano JM, Ridker PM, Brown NJ, Vaughan DE, Moore JH: An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. BMC Bioinformatics 2004, 5:49.
36. Moore JH, Lamb JM, Brown NJ, Vaughan DE: A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. Clin Genet 2002, 62:74-79.

37. Moore JH, Smolkin ME, Lamb JM, Brown NJ, Vaughan DE: The relationship between plasma t-PA and PAI-1 levels is dependent on epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms. Clin Genet 2002, 62:53-59.
38. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol in press.
39. Xu J, Lowey J, Wiklund F, Sun J, Lindmark F, Hsu F, Dimitrov L, Chang B, Turner AR, Liu W, Adami HO, Suh E, Moore JH, Zheng SL, Isaacs WB, Trent JM, Grönberg H: The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk. Cancer Epidemiol Biomarkers Prev 2005, 14:2563-2568.
40. Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS: Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia 2004, 47:549-554.
41. Williams SM, Ritchie MD, Phillips JA 3rd, Dawson E, Prince M, Dzhura E, Willis A, Semenya A, Summar M, White BC, Addy JH, Kpodonu J, Wong LJ, Felder RA, Jose PA, Moore JH: Multilocus analysis of hypertension: a hierarchical approach. Hum Hered 2004, 57:28-38.
42. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001, 69:138-147.
43. Software for MDR and MDR Permutation Testing module. [http://www.epistasis.org].
44. Ritchie MD, Hahn LW, Moore JH: Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003, 24:150-157.
45. Hahn LW, Ritchie MD, Moore JH: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 2003, 19:376-382.
46. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P: Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 2005, 28:171-182.
47. Software for Random Forests. [http://www.stat.berkeley.edu/users/breiman/RandomForests/].

Chapter 3

Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs

A Geert Heidema[1,2,3]
Edith JM Feskens[2]
Pieter AFM Doevendans[4]
Henk JT Ruven[5]
Hans C van Houwelingen[1,6]
Edwin CM Mariman[3]
Jolanda MA Boer[1]

[1] Centre for Nutrition and Health, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands;
[2] Division of Human Nutrition, Wageningen University and Research Centre, Wageningen, The Netherlands;
[3] Department of Human Biology, Maastricht University, Maastricht, The Netherlands;
[4] Heart-Lung Centre Utrecht, University Medical Centre Utrecht, Utrecht, The Netherlands;
[5] Department of Clinical Chemistry, Sint Antonius Hospital, Nieuwegein, The Netherlands;
[6] Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands.

Genetic Epidemiology 2007, 31(8):910–921.

Abstract

Non-parametric approaches have been developed that are able to analyze large numbers of single nucleotide polymorphisms (SNPs) in modest sample sizes. These approaches have different selection features and may not provide similar results when applied to the same dataset. Therefore, we compared the results of three approaches (set association (SAA), random forests (RF) and multifactor dimensionality reduction (MDR)) to select from a total of 93 candidate SNPs a subset of SNPs that are important in determining high-density lipoprotein (HDL)-cholesterol levels.

The study population consisted of a random sample from a Dutch monitoring project for cardiovascular disease risk factors and was dichotomized into cases (low HDL-cholesterol, n=533) and non-cases (high HDL-cholesterol, n=545) based on gender-specific median values for HDL-cholesterol.

Clearly, all three approaches prioritized three SNPs as important (CETP Taq1B, CETP -629 C/A and LPL Ser447X). Two SNPs with weaker main effects were additionally prioritized by RF (ApoC3 3175 G/C and CCR2 Val62Ile), whereas MTHFR 677 C/T was selected in combination with CETP Taq1B as best model by MDR. Obtained p-values for the selected models were significant for SAA ($p=0.0019$), RF ($p<0.01$) and MDR ($p<0.02$).

In conclusion, the application of a combination of multi-locus methods is a useful approach in genetic association studies to select a well-defined set of important SNPs for further statistical and epidemiological interpretation, providing increased confidence and more information compared with the application of only one method.

Introduction

For common traits it is assumed that many different genes, each with allelic variations, contribute to the total observed variability in a trait, with no particular gene having a single large effect [1]. Therefore, studying single nucleotide polymorphisms (SNPs) in genetic association studies is challenging, as it is likely that SNPs do contribute to a certain outcome with only small effects. Logistic regression is the appropriate method to apply to case-control data [2]. However, in genetic association studies where the number of genetic polymorphisms is very large relative to the number of observations, applying logistic regression could easily lead to false positive and false negative results [3]. To overcome this dimensionality problem [4], non-parametric methods have been developed that are able to analyze the effect of large numbers of SNPs in modest sample sizes. These non-parametric methods are referred to as multi-locus methods [5] and have the aim to identify SNPs and interactions between SNPs that are associated with disease or a disease-related endpoint. Multi-locus methods are useful as a first step to select from the large number of measured SNPs a smaller set of most important SNPs and/or combinations between SNPs. For this selected set, logistic regression analysis can be applied in the second step for an epidemiological interpretation.

Selection features of different multi-locus methods and their strength and weaknesses have been discussed [6], including the approaches of set association (SAA) [7], random forests (RF) [8] and multifactor dimensionality reduction (MDR) [9-11]. These non-parametric methods are able to handle the dimensionality problem and have been frequently used in genetic association studies [e.g. 12-17]. Furthermore, software for these methods is freely available and open-source. Although these methods have been applied in several genetic association studies, to our knowledge they have never been used in combination. However, applying methods with different selection features to the same dataset may provide different results. Therefore, we compared in this study the application of SAA, RF and MDR to a real dataset to select from 93 candidate SNPs, a subset of SNPs that are likely to be important in determining HDL-cholesterol levels. After selection of the most important SNPs, we visualized the main and interaction effects for the selected SNPs by interaction graphs for further statistical interpretation. Logistic regression analysis was performed to model important main and interaction effects for an epidemiological interpretation.

Methods

Study population

The study population consisted of a random sample of 1,078 individuals from a Dutch monitoring project for cardiovascular disease risk factors [18]. This project was carried out between 1987 and 1991 among 35,488 men and women. Participants were between 20 and 59 years old. Non-fasting blood samples were obtained at the Municipal Health Services in three Dutch municipalities (Amsterdam, Doetinchem and Maastricht) and stored in EDTA-coated vacutainer tubes. All participants gave written informed consent and approval was

obtained from the medical ethics committees of Leiden University and the Netherlands Organisation for Applied Scientific Research.

The random sample of 1,078 individuals was divided into subjects with low (cases) and high (non-cases) HDL-cholesterol levels according to gender-specific median HDL-cholesterol levels. Median values were 1.11 and 1.35mmol/l for men and women, respectively. Men and women with HDL-cholesterol levels below the median were designated as cases (n=533), whereas those above the median were designated as non-cases (n=545). Characteristics of the study population are shown in Table 1.

Table 1: Characteristics of the study population.

Study population	Men N=506	Women N=572	Total N=1078
Age (years)	42.0 ± 10.7	41.5 ± 10.9	41.8 ± 10.8
HDL-cholesterol			
Median (mmol/l)	1.11	1.35	1.24
Low (n)	250	283	533
High (n)	256	289	545
Total cholesterol (mmol/l)	5.55 ± 1.11	5.48 ± 1.17	5.52 ± 1.14
BMI (kg/m^2)	24.8 ± 3.35	24.6 ± 3.97	24.7 ± 3.69

Results are presented as mean ± SD unless otherwise stated.
BMI: Body mass index

Laboratory analyses

HDL-cholesterol

HDL-cholesterol was determined enzymatically using a Boehringer test kit (Monotests Cholesterol High Performance, Boehringer Mannheim GmbH, Germany) within 3 weeks after storage, after precipitation of ApoB containing lipoproteins with magnesium phosphotungstate [19, 20].

Genotyping

Genomic DNA was extracted from frozen buffy coats as described by Hoebee et al. [21]. A number of 117 SNPs have been genotyped by the RMS research assay for CVD genetics and the RMS research assay for inflammatory disease genetics (Roche Molecular Systems, Inc., Alameda, CA) [22]. In addition, nine SNPs have been determined by pyrosequencing or restriction fragment-length polymorphism analysis (details are available upon request). Genotyping was performed without knowledge about the case status of the samples.

Data preparation

Of the 126 SNPs that were determined (see supplemental table), 16 SNPs were genotyped by two different methods. For these SNPs, genotypes concorded between methods for 92.8–99.7% and redundant SNPs were discarded from further analyses (based on the number of missing values or deviation from Hardy-Weinberg equilibrium). Additionally, those SNPs that clearly deviated from Hardy-Weinberg equilibrium ($p<0.01$) using a χ^2-test, were removed from the dataset (n=10). One SNP was removed from the dataset because many subjects had missing values (n=357), whereas five SNPs with minor allele frequencies lower than 0.01 were also discarded for the analyses. Finally, we removed one of two SNPs in the LTA-gene that were highly correlated ($r^2>0.85$), as these SNPs would otherwise become interchangeable in the analyses. This resulted in a total of 93 SNPs available for analyses. For SNPs containing less than five observations for the homozygous mutant genotype (in cases and/or non-cases), homozygous mutant individuals were pooled with heterozygous individuals. Missing values were imputed with the value of the most frequent class before performing analyses with the different methods.

Statistical analyses

Fisher's exact test

To compare the results of the different methods with the results of univariate analyses, the Fisher's exact test was performed. We used SAS software version 9.1 (SAS institute, Inc., Cary, NC) to study univariate associations between SNPs and HDL-cholesterol group.

Set association approach

SAA has been developed by Hoh et al. [7]. SAA prioritizes SNPs by a univariate test-statistic. Sums of test-statistics are computed, starting with the SNP that has the highest univariate test-statistic. Each time the SNP with the next highest test-statistic is added to the sum. Thereby, sums of test-statistics consisting of increasing numbers of SNPs are formed. Permutation tests are performed to obtain p-values for the different sums. SNPs in the sum with the smallest p-value are selected and another permutation test is performed to obtain an overall p-value.

In SAA different test-statistics can be used. We applied both the χ^2 of genotype by HDL-cholesterol group and the χ^2 of allele by HDL-cholesterol group as test-statistic. The number of permutated datasets used was equal to 10,000. The software we used for SAA is freely available at http://www.genemapping.cn/sumstat.html.

Random forests approach

RF was developed by Breiman [8]. Application of classification RF to SNP data has been described previously [16, 23]. In RF, an ensemble of tree models is used to predict case status. Each tree recursively splits the total dataset into smaller and more homogeneous subgroups of cases and non-cases, whereby the total sample for each tree is obtained by bootstrap sampling. With bootstrap sampling, sampling is performed with replacement and some individuals are sampled more than once while others are left out. The sampled observations are used to construct the tree, whereas a prediction is obtained for each left-out

individual. Aggregating the predictions over the different trees in which the individual was left-out, a prediction for this individual is obtained for the ensemble of trees, which is called the forest. The proportion of misclassified cases and non-cases provides the prediction error of the forest. Another important feature is that the predictor that gives the best partitioning in cases and non-cases at a certain split is not selected from the total number of predictors but from a smaller random sample of predictors. This parameter is referred to as m_{try}. The default value for m_{try} is the square root of the number of variables to be analyzed in the dataset (in this dataset $\sqrt{total\ nr\ of\ SNPs} = 9$). We varied the value of m_{try} from 5–20 to determine the value of m_{try} resulting in the lowest prediction error. The lowest prediction error was obtained with a m_{try} value of 5 and results will therefore be shown for this number.

RF provides a measure of importance for each SNP, referred to as the importance index, by comparing the predictive performance of the forest for all SNPs with the predictive performance of the forest for all SNPs but with the values for one SNP randomly permuted for the left-out individuals. Larger differences in the predictive performance indicate more important SNPs. RF does not provide a threshold to define SNPs as being important. To define such a threshold we performed several analyses to examine whether there was a subset of SNPs that was consistently prioritized. Using the m_{try} value of 5, a threshold choice to select SNPs as important was obtained by performing the analyses several times, each time using a different seed value (the seed value controls the random number generator). The threshold to select SNPs was defined by the value of importance index I_m, where the ranking of SNPs started to deviate between different analyses. SNPs with an importance index higher than this threshold were selected.

To validate whether the RF model does significantly predict case status, a permutation test [24, 25] was performed to obtain the significance of the prediction error. A 100 datasets were formed with labels of case status permuted randomly. By applying RF to each permuted dataset, a prediction error was obtained for each RF model, thereby forming the distribution of the prediction error under no association. The number of permuted datasets that had a prediction error equal or lower than the prediction error for the observed dataset provided the significance of the prediction error. A significant result indicates that significant associations are present in the dataset.

For each of the different analyses the number of trees in the forest was set to 30,000. For the analyses we used the R package randomForest written by [26, 27], which is based on the original FORTRAN code from Breiman et al. (freely available at www.stat.berkeley.edu/users/breiman/randomforests/). This R package is freely available from the CRAN website (http://cran.r-project.org/).

Multifactor dimensionality reduction method

MDR was developed by Ritchie et al. [12] and has been described in detail in several reviews [9-11]. MDR is a non-parametric data mining approach that uses constructive induction or attribute construction to reduce two or more SNPs, for example, to a new single variable that is then evaluated using a classifier such as naïve Bayes or logistic regression. In MDR, each multi-locus genotype of a SNP combination is assigned to a high-risk or low-risk group, depending on the ratio of cases and non-cases with this multi-locus genotype. If this ratio exceeds a certain threshold, this multi-locus genotype is assigned to

as high-risk, otherwise it is assigned to as low-risk. In this study, the ratio of cases to non-cases present in the dataset was used as threshold. By assigning all multi-locus genotypes for a certain combination of SNPs to either high-risk or low-risk, MDR reduces the number of multi-locus genotypes to one risk factor consisting of two levels, high-risk or low-risk. The aim is to construct a new risk factor that facilitates the detection of non-linear interactions among SNPs such that the prediction of the outcome variable is improved over the original representation of the data. In the MDR software the performance of a newly constructed risk factor to classify and predict case status is evaluated by multi-fold cross-validation and permutation testing.

We applied the MDR software (freely available at http://www.epistasis.org), using 10-fold cross-validation, to determine for our dataset the best model for main effects, two-SNP and three-SNP combinations, yielding one best model for each. The 10-fold cross-validation was repeated 10 times, using each time a different seed value, to protect against chance divisions of the dataset. Of the three best models, the model with the highest average testing accuracy and cross-validation consistency was denoted as final best model. The cross-validation consistency is the number of times a model is selected as best model among the validation sets and thereby indicates the importance of a model. A model containing important SNPs will be selected regardless of the division of the data for cross-validation. For each of the best models with a cross-validation consistency less than 100, we performed a forced MDR analysis to obtain an unbiased estimate of the testing accuracy. Finally, applying the MDR permutation module we tested the significance of the testing accuracy of the final best model to validate whether this model was significantly associated with case status. A 100 datasets with case status permuted randomly were formed and MDR was applied to each of these datasets. For each permuted dataset the best model was determined. The proportion of testing accuracies of the permuted samples that equalled or exceeded the testing accuracy of the observed best model provided the significance of the final best model.

Interaction graphs

For further statistical interpretation, we visualized the importance of the relationships between the SNPs selected by the different methods in two interaction graphs [11, 28, 29]. First an entropy-based interaction graph is used to compare the independent main effects of polymorphisms with the pair-wise interaction effect between polymorphisms. Positive or negative entropy values for pair-wise SNP interactions indicate that besides the two individual SNP effects additional information is explained by their interaction. Positive entropy values indicate a synergistic interaction, whereas negative entropy values indicate the presence of redundant information. To create the entropy-based interaction graph the Orange machine learning software was used, which is freely available and open source [30] at http://www.ailab.si/orange. Second, we used an interaction dendrogram to visualize the importance of the relationships between the SNPs. The interaction dendrogram is constructed using hierarchical cluster analysis and is implemented as a feature in the MDR software. The graph displays the interactions between SNPs whereby strongly interacting SNPs are located closely together at the leaves of the tree. Furthermore, information is provided whether interactions between SNPs are synergistic or that redundancy is present.

Logistic regression analysis

Logistic regression analysis has been performed for the important main and interaction effects obtained with the multi-locus methods and the interaction graphs. All relevant SNPs and SNP combinations were included in one model. For each SNP, wild-type homozygous individuals are used as the reference group in the analyses and two dummy variables have been created, one for the heterozygous and one for the homozygous mutant individuals. The number of homozygous mutant individuals with the ApoC3 3175 G/C polymorphism was low, therefore these individuals have been pooled with the heterozygous individuals before performing the analyses. SAS software version 9.1 (SAS institute Inc., Cary, NC) was used to perform the logistic regression analyses.

Results

Fisher's Exact test

For each SNP in the dataset, the distribution of cases and non-cases over the different genotypes and the corresponding significance obtained by Fisher's exact test are shown in the supplemental table (see supplemental data). About ten SNPs were significant at the 0.05 level.

Set association approach

Using χ^2 of genotype by HDL-cholesterol group as test-statistic, p-values for the sums of test-statistics were obtained by permutation testing. Clearly, p-values largely decreased for the three SNPs with the highest test-statistics (CETP Taq1B, CETP -629 C/A and LPL Ser447X) included in the sum (see figure 1, results are shown for the first 30 SNPs). Adding more SNPs to the sum did not result in a further decrease of the p-value. The sum including the first three SNPs resulted in the lowest p-value and therefore the first three SNPs were selected as important. The overall p-value for this sum was equal to p=0.0019. When these three SNPs with the highest test-statistic were excluded, a non-significant result was obtained for the remaining SNPs (p=0.55).

For the χ^2 of allele by HDL-cholesterol group test-statistic, the sum containing the three SNPs with the highest test-statistics (CETP Taq1B, CETP -629 C/A and LPL Ser447X) also resulted in the lowest p-value (results not shown). The overall p-value was highly significant (p=0.0005) and by excluding the first three SNPs a non-significant result was obtained (p=0.40). Thus application of the test-statistic of χ^2 of allele by HDL-cholesterol group resulted in selection of the same SNPs as with the test-statistic of χ^2 of genotype by HDL-cholesterol group.

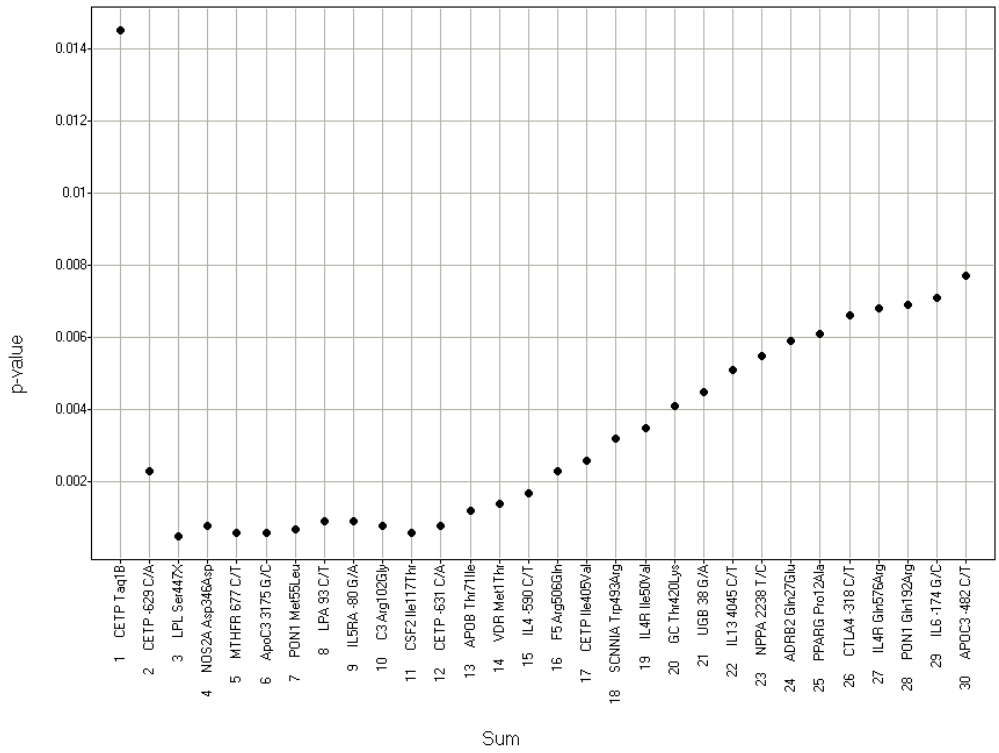


Figure 1: Results of SAA using χ^2 of genotype by HDL-cholesterol group as test-statistic. P-values for the sums of test-statistics are shown. Values at the x-axis correspond to the number of single nucleotide polymorphisms included in the sum.

Random Forests

For m_{try} value 5, SNPs were ranked according to their importance index for different seed values. Results are shown for seed values 17 and 43 in figure 2, where the importance indices of the SNPs are plotted against each other. Ranking of SNPs was similar for the five SNPs with the highest importance indices and started to deviate for lower ranked SNPs. Different m_{try} values consistently resulted in prioritization of the same 5 SNPs (results not shown). Therefore, these five SNPs (CETP Taq1B, CETP -629 C/A, LPL Ser447X, ApoC3 3175 G/C and CCR2 Val62Ile) were selected as important.

The prediction error for the observed dataset was equal to 45%. None of the 100 permuted datasets had a lower prediction error compared with the observed prediction error. Therefore, the significance of overall prediction error rate of the forest is equal to $p<0.01$, indicating that significant SNP effects are present in this dataset.

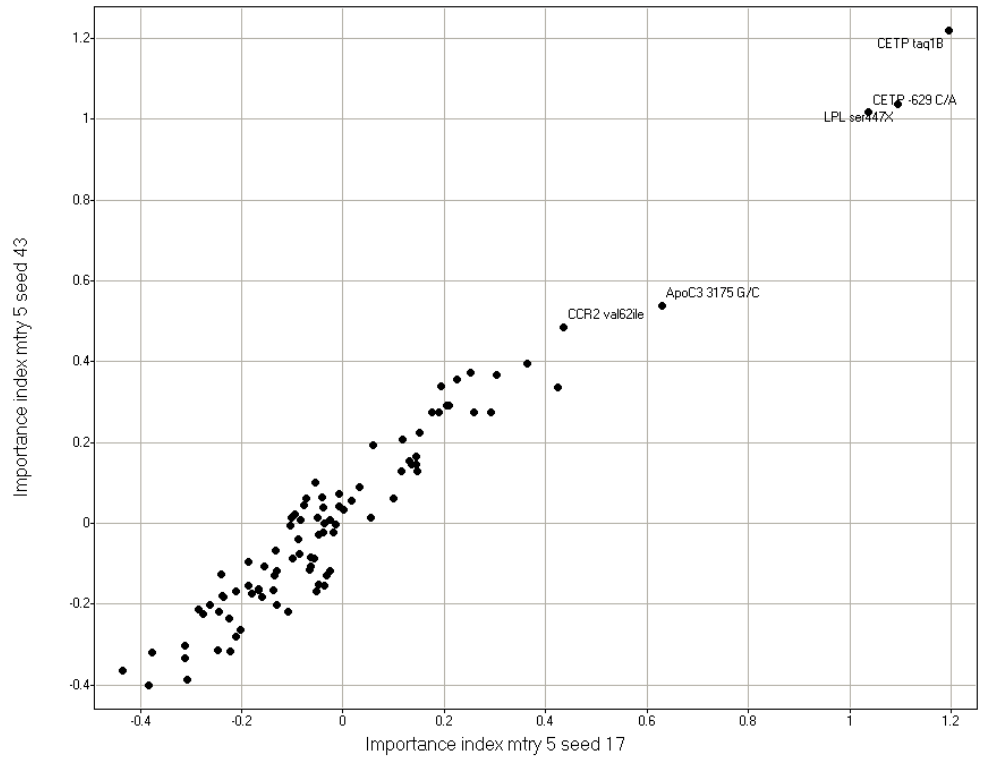


Figure 2: Importance indices of the single nucleotide polymorphisms obtained by RF. Important indices obtained by m_{try} value 5 using seed 17 are plotted against the importance indices obtained by m_{try} value 5 using seed 43.

Multifactor Dimensionality Reduction method

Prioritization of main effects and two-SNP combinations by their classification accuracy shows that the four main effects and the four 2-SNP combinations with the highest classification accuracy consist of the SNPs LPL Ser447X, CETP Taq1B, CETP -629 C/A and MTHFR 677 C/T (see table 2).

For the main effects and the two-SNP and three-SNP combinations, the best models are shown in table 3. The best two-SNP and three-SNP combinations both had the highest average testing accuracy. Therefore, to obtain the final best model we compared the cross-validation consistency of the two-SNP model with that of the three-SNP model. The model containing two SNPs (MTHFR 677 C/T and CETP Taq1B) had the highest cross-validation consistency and was therefore selected as final best model. Frequencies of cases and non-cases for the different multi-locus genotypes for this model are shown in figure 3. Using a permutation test, the best model obtained with MDR was found to be significantly associated with case status ($p<0.02$).

Table 2: Prioritization by MDR of SNPs and combinations of two SNPs by classification accuracy.

SNP main effect		Average classification accuracy
LPL Ser447X		54.7
CETP Taq1B		54.6
CETP -629 C/A		54.5
MTHFR 677 C/T		54.4
PON1 met55leu		53.8
IL5RA 80 G/A		53.6
NOS2A asp346asp		53.4
ApoC3 3175 G/C		53.3
2 SNP-combination		
CETP Taq1B	MTHFR 677 C/T	57.9
CETP -629 C/A	LPL Ser447X	57.6
CETP Taq1B	LPL Ser447X	57.4
CETP -629 C/A	MTHFR 677 C/T	57.2
CETP Taq1B	IL5RA 80 G/A	57.0
CETP Taq1B	PON1 met55leu	56.9
CETP -629 C/A	IL5RA 80 G/A	56.8
CETP -629 C/A	IL9 thr113met	56.7

Table 3: Best MDR models for the main effects, two-SNP and three-SNP combinations.

Model	Average testing accuracy	Cross-validation consistency
LPL Ser447X	54.3	34/100
MTHFR 677 C/T CETP Taq1B	**57.9**	**62/100**
MTHFR 677 C/T CETP Taq1B C3 Arg102Gly	57.9	31/100

Values for the final best model are in bold.

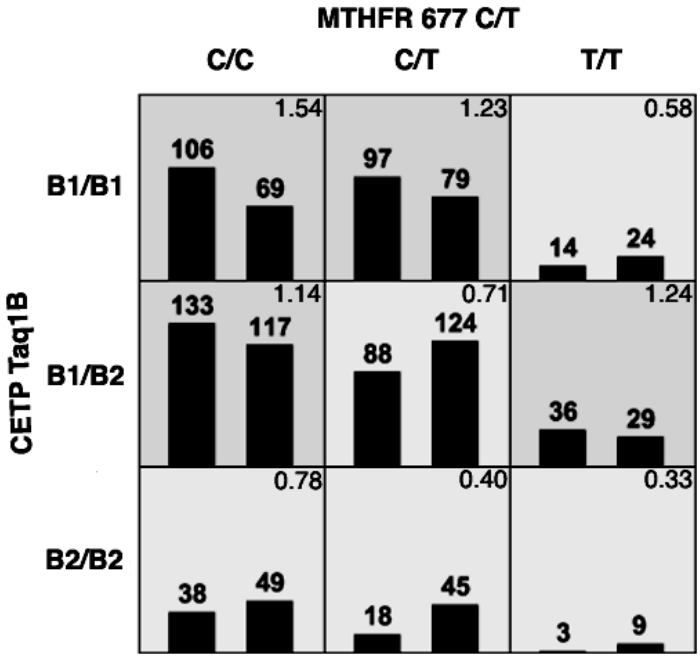


Figure 3: Frequencies of cases and non-cases for the different multi-locus genotypes of the best model obtained by MDR (combination of MTHFR 677 C/T and CETP Taq1B). For each cell the first column represents the number of cases, the second column the number of non-cases. Dark gray cells indicate high-risk multi-locus genotypes, light gray cells indicate low risk multi-locus genotypes. For each multi-locus genotype the ratio of cases to non-cases is depicted at the upper right corner.

Overall results

A summary of the results from the different multi-locus methods and the Fisher's exact test is given in table 4. SNPs with clear main effects (CETP Taq1B, CETP -629 C/A, LPL Ser447X) are prioritized by all three multi-locus methods. Additional SNPs are prioritized by RF and MDR. RF additionally prioritized ApoC3 3175 G/C and CCR2 Val62Ile, whereas MDR additionally selected MTHFR 677 C/T in combination with CETP Taq1B in the final best model. The second best two-SNP combination obtained by MDR consisted of two SNPs (CETP -629 C/A and LPL Ser447X) that were also selected by SAA and RF. Combining these results, the SNPs prioritized by the different methods are CETP Taq1B, CETP -629 C/A, LPL Ser447X, MTHFR 677 C/T, ApoC3 3175 G/C and CCR2 Val62Ile.

Table 4: Prioritization and selection of SNPs based on Fisher's Exact test (p<0.05), SAA, RF and MDR.

SNP	Fisher's Exact test	SAA	RF	MDR
	p-value	χ^2-genotype p-value	Importance index (seed 17)	
CETP Taq1B	0.0002	0.0145	1.20	1*
CETP -629 C/A	0.0005	0.0023	1.04	2†
LPL Ser447X	0.0008	0.0005	1.10	2†
ApoC3 3175 G/C	0.007		0.63	
NOS2A Asp346Asp	0.01			
MTHFR 677 C/T	0.01			1*
CETP -631 C/A	0.03			
PON1 Met55Leu	0.04			
LPA 93 C/T	0.04			
IL5RA -80 G/A	0.05			
CCR2 Val62Ile	0.15		0.43	
Significance of the model		0.0019	<0.01	<0.02‡

* Selected as best two-SNP combination
† Second-best two-SNP combination
‡ Significance of the best two-SNP combination

Interaction graphs

The interaction entropy graph for the six SNPs selected by the different methods is shown in figure 4. Although the entropy removed by the different SNPs and interactions between SNPs is small, the interaction graph confirms the results obtained in this study. CETP Taq1B, CETP -629 C/A and LPL Ser447X remove the most entropy (1.17, 1.02 and 0.91%, respectively), which corresponds with the Fisher's exact test and the prioritization of these SNPs by all three multi-locus methods. Also, small synergistic interaction effects between the CETP polymorphisms and MTHFR 677 C/T are found. These interactions remove about the same amount of entropy (0.74 and 0.71%). This corresponds with the results of MDR that selected the combination of MTHFR 677 C/T and CETP Taq1B as final best model. Together MTHFR 677 C/T and CETP Taq1B explain 2.45% (1.17+0.57+0.71) of the entropy in HDL-cholesterol group.

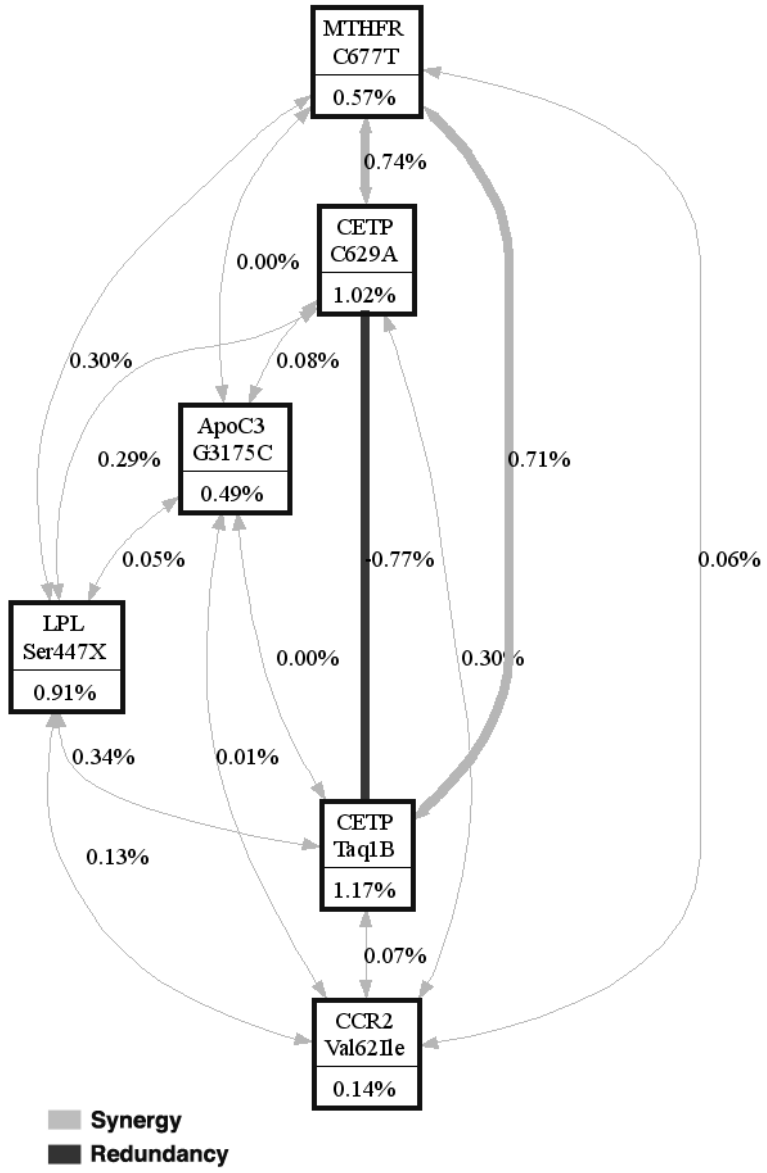


Figure 4: Entropy-based interaction graph. The percentages of entropy of HDL-cholesterol group explained by the different SNPs are shown in the boxes. The numbers by the arrows correspond to the percentages of entropy of HDL-cholesterol group explained by the two-way interactions between single nucleotide polymorphisms. **For full color figure, see page 164.**

The negative entropy value (-0.77%) between the two CETP polymorphisms indicates redundant information that is present in both polymorphisms. The redundancy is an indication of the presence of linkage disequilibrium between these SNPs. The effect of ApoC3 3175 G/C is small and ApoC3 3175 G/C does not interact with any of the other SNPs. Effects including the CCR2 Val62Ile polymorphism are almost absent.

The interaction dendrogram obtained with the MDR software is shown in figure 5. Redundancy between the CETP polymorphisms is clearly shown in the dendrogram. Also, CCR2 Val62Ile is on a separate branch from the other SNPs, indicating that CCR2 Val62Ile does not interact with any of the other SNPs. These results correspond to the results obtained with the interaction entropy graph. However, we also observed differences between the two interaction graphs. In the dendrogram, the interactions between the CETP polymorphisms and MTHFR 677 C/T are not clearly present. Also, the dendrogram shows redundancy between ApoC3 3175 G/C and the CETP polymorphisms, and between LPL Ser447X and MTHFR 677 C/T, which was not found by the interaction entropy graph.

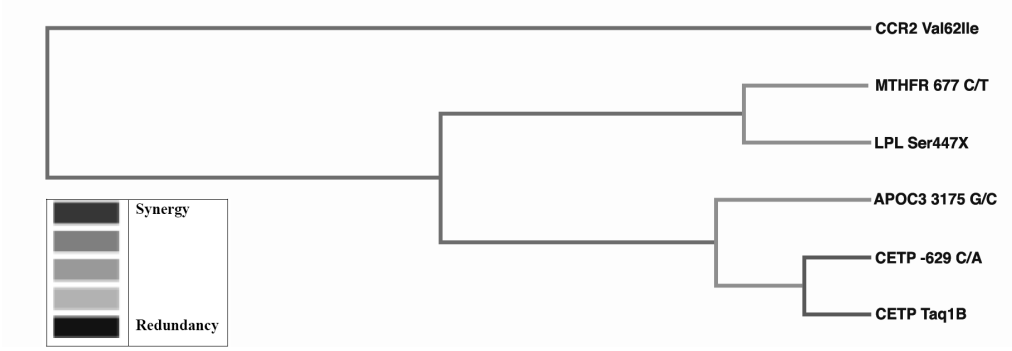


Figure 5: Interaction dendrogram. Stronger interactions between single nucleotide polymorphisms are visualized by depicting SNPs more closely together at the leaves of the tree (right side of the graph). **For full color figure, see page 165.**

Logistic regression analyses

Based on the findings of the multi-locus methods and the interaction graphs, logistic regression analysis was performed for further epidemiological interpretation. Of the six prioritized SNPs, four were included in the model either as main effect (LPL Ser447X and ApoC3 3175 G/C) or multi-locus genotype effect (for the combination of MTHFR 677 C/T and CETP Taq1B). The interaction between MTHFR 677 C/T and CETP -629 C/A was left out of the model because of the redundancy between the CETP polymorphisms. Furthermore, CCR2 Ile62Val was left out because of its low entropy value and the absence of pair-wise interactions including this SNP.

Risk estimates for the main and interaction effects are shown in table 5. Having at least one mutant allele for the LPL Ser447X polymorphism results in a significant decrease ($p<0.01$) in risk of low HDL-cholesterol (odds ratio (OR)=0.55). A significant increase in risk ($p<0.01$) was obtained for the ApoC3 3175 G/C polymorphism (OR=1.51). For the combination of MTHFR 677 C/T and CETP Taq1B significant multi-locus genotype effects

were found. All OR for the multi-locus genotypes are lower than 1, indicating that homozygous wildtype genotypes for both polymorphisms give the highest risk for low HDL-cholesterol.

Table 5: Odds ratios for low HDL-cholesterol according to LPL Ser447X, ApoC3 3175 G/C and the combination of MTHFR 677 C/T with CETP Taq1B.

Effect	Odds ratio	95% Confidence intervals	
LPL Ser447X C/C	1.00 (ref)		
LPL Ser447X C/G G/G	0.55**	0.41	0.74
Apo C3 3175 G/G	1.00 (ref)		
Apo C3 3175 G/C C/C	1.51**	1.11	2.07
MTHFR 677 C/C – CETP Taq1B B1/B1	1.00 (ref)		
MTHFR 677 C/C – CETP Taq1B B1/B2	0.73	0.49	1.08
MTHFR 677 C/C – CETP Taq1B B2/B2	0.47**	0.28	0.80
MTHFR 677 C/T – CETP Taq1B B1/B1	0.78	0.51	1.19
MTHFR 677 C/T – CETP Taq1B B1/B2	0.45**	0.30	0.69
MTHFR 677 C/T – CETP Taq1B B2/B2	0.25**	0.14	0.48
MTHFR 677 T/T – CETP Taq1B B1/B1	0.35**	0.17	0.73
MTHFR 677 T/T – CETP Taq1B B1/B2	0.75	0.42	1.35
MTHFR 677 T/T – CETP Taq1B B2/B2	0.19*	0.05	0.73

Odds ratios were obtained by including all effects into one multivariate logistic regression model.
Ref: reference category
* Significant at the 0.05 level
** Significant at the 0.01 level

Discussion

In this study, we have applied three different multi-locus methods that have been developed for analyzing large numbers of SNPs in relation to disease or a disease-related endpoint. The findings in this study show that applying different multi-locus methods has several advantages compared with applying only one method. First, because SNPs with clear main effects are prioritized by all the different methods, the confidence is strengthened that these SNPs are relevant for the endpoint of interest. A second advantage is that by applying different multi-locus methods more information is obtained than applying only one method. Different SNPs were additionally prioritized by RF and MDR. Furthermore, univariate analyses using Fisher's exact test resulted in ten SNPs to be significantly associated with HDL-cholesterol group at the significance level of $p<0.05$. By applying multi-locus methods a smaller set of six SNPs was selected and the chance to falsely select SNPs as important was thereby likely reduced. Thus, in genetic association studies, the application of different multi-locus methods to prioritize and select SNPs is a useful approach to identify a well-defined set of SNPs that is likely to be important.

The three different approaches all yielded significant results, which indicates that significant associations are present in this dataset. In MDR, only one best model is selected. In our study the combination of MTHFR 677 C/T and CETP Taq1B best predicted low HDL-cholesterol levels. However, the second best two-SNP combination, consisting of CETP -629 C/A and LPL Ser447X, only had a slightly lower classification performance

than the best model (57.6 versus 57.9, respectively). These SNPs were also selected by SAA and RF and are therefore likely to have an important contribution to HDL-cholesterol levels. By selecting only one best model to predict HDL-cholesterol levels, MDR leaves out other SNPs and/or combinations of SNPs that can be important as well. Nevertheless, modeling combinations of SNPs, MDR was the only method that was able to capture MTHFR 677 C/T as important in combination with CETP Taq1B. Visualizing pair-wise interactions by the interaction entropy graph confirmed that this combination explained the largest amount in entropy of HDL-cholesterol group. Thus, comparing the results of the different methods is useful, as more information can be obtained compared with applying only one method.

SAA and MDR are developed for analyzing dichotomous endpoints. Therefore, to be able to compare the different multi-locus methods we dichotomized HDL-cholesterol levels into high and low-HDL-cholesterol based on gender-specific median values. However, we realize that by dichotomizing HDL-cholesterol levels not all information is used. RF is the only method that can also be applied to continuous endpoints, using regression trees instead of classification trees. Therefore, we also performed analyses with RF using regression trees to select SNPs that are important in determining continuous HDL-cholesterol levels. Applying RF using regression trees resulted also in the prioritization of CETP Taq1B, CETP -629 C/A, LPL Ser447X, MTHFR 677 C/T and ApoC3 3175 G/C. However, CCR2 Val62Ile was not prioritized (results not shown). Adding sex to the RF model slightly changed the prioritization of SNPs. Sex was highly prioritized and SNPs prioritized included CETP Taq1B, CETP -629 C/A, LPL Ser447X and MTHFR 677 C/T. ApoC3 3175 G/C was less highly prioritized and again, CCR2 Val62Ile was not prioritized. Thus, the results of RF using regression trees replicate the finding of MDR that MTHFR 677 C/T is also of importance, whereas the importance of CCR2 Val62Ile was not replicated. Both the entropy-based interaction graph and interaction dendrogram showed no important pair-wise effects for CCR2 Val62Ile. Additionally, the entropy value for CCR2 Val62Ile was very low, indicating that this SNP may not be important. Thus, whether CCR2 Val62Ile is relevant in determining HDL-cholesterol levels is questionable and may therefore be a chance finding. However, RF is the only method that is able to handle genetic heterogeneity and in RF the importance of each SNP is obtained in the context of all other SNPs. Therefore, it may also be that CCR2 Val62Ile is of importance in only a subgroup or that this SNP is important in interactions containing more than two SNPs.

In theory, it could be possible that all methods modeled the same noise in the dataset. However, besides statistical interpretation, biological interpretation of the results is also an important tool to verify whether selected SNPs are relevant. Of all the 93 SNPs that are present in the dataset, only a smaller subset consists of polymorphisms that are located in candidate genes of lipid metabolism. Associations between the SNPs selected by the different methods (with the exception of the CCR2 polymorphism) and HDL-cholesterol levels have been found in previous studies [31, 32]. Thus, the biological relevance of the selected SNPs also increases the confidence that the selected SNPs in this study are relevant for HDL-cholesterol levels. Although an association between MTHFR 677 C/T and HDL-cholesterol has been found previously, this association may be indirect. It is suggested that the MTHFR polymorphism has an effect on homocysteine levels and that increased homocysteine levels lead to a decrease in levels of HDL-cholesterol [31, 33].

How to define a threshold for the selection of SNPs by RF is not straightforward and needs to be further examined. Besides the approach used in this study, we also examined another approach to define a threshold. In this second approach, we performed a permutation test to obtain the significance of the importance index for the different SNPs. However, this approach did not yield a clear threshold (results not shown) and was therefore not useful to define a threshold in this dataset. Besides comparing different analyses within RF, biological verification may also be useful to denote a threshold to select SNPs as being important using RF. An example of using biological relevance to define a threshold for the selection of prioritized variables is shown by Enot et al. [34]. However, biological relevance is not always known beforehand and relying solely on biological information may lead to discarding polymorphisms in genes that have a relationship with the outcome of interest that has not been found previously in the literature.

Genotype discordancies were found for SNPs that have been genotyped by two different methods. For this study genotype discordancies did not influence the comparison of results of the different methods as all methods have been applied to the same dataset. Still, it may have affected the biological interpretation of the results. However, the discordancies were found to be random. Furthermore, the performance of MDR to detect risk SNPs has shown to be excellent in the presence of 5% genotyping errors [35]. Therefore, we do not think that genotype discordancies will have had a large influence on the biological interpretation of the results.

In this study, we applied a two-step approach. In the first step, three multi-locus methods were combined to prioritize and select SNPs related to HDL-cholesterol group. In the second step, the set of selected SNPs was used for further statistical and epidemiological interpretation. The multi-locus methods applied in this study can also be used within a multi-step approach, suggested by Moore et al. [11]. In this multi-step approach SNPs are also prioritized and selected in the first step. Prioritization and selection can be performed by univariate tests, SAA and/or RF, but other measures of importance can be used as well. The advantage of the RF approach is that it can handle genetic heterogeneity in the dataset, if present [23, 36]. However, in the multi-step approach of Moore et al. MDR is not used to select SNPs, but the constructive induction algorithm of MDR is used in the second step to construct new variables for combinations of SNPs. By constructing new variables MDR reduces for each SNP combination the number of multi-locus genotypes to one variable consisting of two risk groups (high and low risk). These newly constructed variables can then be evaluated in the third step by machine learning approaches. At this step, multi-fold cross-validation can be used, but also RF [37]. If the number of observations is large enough relative to the number of effects selected, logistic regression can also be used for an epidemiological interpretation. After evaluation of the constructed variables, statistical interpretation of the results can be performed using interaction graphs. Irrespective of the approach applied (two-step, multi-step), inferring the biological relevance of the selected SNPs is an important last step. For biological interpretation of the selected SNPs and/or combinations between SNPs, pathway analyses can be performed (e.g., by GenMAPP, freely available at http://www.genmapp.org). Also, systems biology is proposed [38] to come to an understanding of how statistical interactions relate to biological interactions.

Conclusion

In this study, three different multi-locus methods have been applied to prioritize and select SNPs that are related to HDL-cholesterol. Combining the results of these methods strengthens the confidence of the relevance of similarly selected SNPs and more information can be obtained as different SNPs are additionally selected by different methods. Therefore, the application of different multi-locus methods is a useful approach in genetic association studies to select a well-defined set of important SNPs for further statistical and epidemiological interpretation.

Acknowledgements

This project has been carried out in the framework of the Centre for Human Nutrigenomics. The Monitoring Project on Cardiovascular Disease Risk Factors was financially supported by the Ministry of Public Health, Welfare and Sports of the Netherlands. We thank the epidemiologists and fieldworkers of the Municipal Health Services in Amsterdam, Doetinchem and Maastricht for their contribution to the data collection, A. Blokstra for data management and logistic support and Prof. Dr. Ir. D. Kromhout for project management. The data collection of this study was partly supported by grant 98B067 from the Netherlands Heart Foundation. We thank P. van Impelen, H. Hodemaekers and H. Alpar for their contribution to the laboratory analyses performed in this study and Roche Molecular Systems for providing their prototype CVD and Inflammatory Marker genotyping assays. We also thank Dr. T. Travis for allowing the use of the computer cluster at the Rowett Research Institute.

References

1. Risch J: Searching for genetic determinants in the new millennium. Nature 2000, 405:847–856.
2. Le Cessie S, Van Houwelingen JC: Ridge estimators in logistic-regression. Appl Stat J Stat Soc Series C 1992, Vol 41 1:191–201.
3. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein, AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996, 49: 1373–1379.
4. Bellman R: Adaptive Control Processes. Princeton NJ: Princeton University Press; 1961.
5. Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003, 4: 701–709.
6. Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, Van der AD, Feskens EJM: The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genet 2006, 7:23.
7. Hoh J, Wille A, Ott J: Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 2001, 11:2115–2119.
8. Breiman L: Random forests. Machine learn 2001, 45:5–32.
9. Moore JH: Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 2004, 4:795–803.
10. Wilke RA, Reif DM, Moore JH: Combinatorial pharmacogenetics. Nat Rev Drug Discov 2005, 4:911–918.

11. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 2006, 241:252–261.
12. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001, 69:138–147.
13. Zee RY, Hoh J, Cheng S, Reynolds R, Grow MA, Silbergleit A, Walker K, Steiner L, Zangenberg G, Fernandez-Ortiz A, Macaya C, Pintor E, Fernandez-Cruz A, Ott J, Lindpaintner K: Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. Pharmacogenomics J 2002, 2:197–201.
14. Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS: Multifactor dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia 2004, 47:549–554.
15. De Quervain DJ, Poirier R, Wollmer MA, Grimaldi LM, Tsolaki M, Streffer JR, Hock C, Nitsch RM, Mohajeri MH, Papassotiropoulos A: Glucocorticoid-related genetic susceptibility for Alzheimer's disease. Hum Mol Genet 2004, 13:47–52.
16. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P: Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 2005, 28:171–182.
17. Xu J, Lowey J, Wiklund F, Sun J, Lindmark F, Hsu F, Dimitrov L, Chang B, Turner AR, Liu W, Adami HO, Suh E, Moore JH, Zheng SL, Isaacs WB, Trent JM, Grönberg H: The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk. Cancer Epidemiol Biomarkers Prev 2005, 14:2563–2568.
18. Verschuren WMM, Boerma GJ, Kromhout D: Total and HDL-cholesterol in the Netherlands: 1987-1992. Levels and changes over time in relation to age, gender and educational level. Int J Epidemiol 1994, 23:948–956.
19. Lopes-Virella MF, Stone P, Ellis S, Colwell J: Cholesterol determination in high-density lipoproteins separated by three different methods. Clin Chem 1977, 23:882–884.
20. Kattermann R, Jaworek D, Mo¨ ller G, Assmann G, Bjo¨rkhem I, Svensson L, Borner K, Boerma G, Leijnse B, Desager JP, Harwent C, Kupke I, Trinder P: Multicentre study of a new enzymatic method of cholesterol determination. J clin chem. Clin Biochem 1984, 22:245–251.
21. Hoebee B, Rietveld E, Bont L, Oosten M, Hodemaekers HM, Nagelkerke NJ, Neijens HJ, Kimpen JL, Kimman TG: Association of severe respiratory syncytial virus bronchiolitisv with interleukin-4 and interleukin-4 receptor alpha polymorphisms. J Infect Dis 2003, 187:2–11.
22. Hoppe C, KlitzW, Cheng S, Apple R, Steiner L, Robles L, Girard T, Vichinsky E, Styles L: Gene interactions and stroke risk in children with sickle cell anemia. Am Soc Hematol 2004, 4: 701–709.
23. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004, 5:32.
24. Cox DR, Hinkley DV: Theoretical Statistics. London: Chapman and Hall; 1974.
25. Lyons-Weiler J, Pelikan R, Zeh HJ, Whitcomb DC, Malehorn DE, Bigbee WL, Hauskrecht M: Assessing the statistical significance of the achieved classification error of classifiers constructed using serum peptide profiles, and a prescription for random sampling repeated studies for massive high-throughput genomic and proteomic studies. Cancer Informatics 2005, 1:53–77.
26. Liaw A,Wiener M: Classification and regression by random-Forest. Rnews 2002, 2:18–22.
27. R Development Core Team: R: a language and environment for statistical computing. [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna, 2004.
28. Jakulin A, Bratko I: Analyzing attribute dependencies. Lect Notes Artif Intell 2003, 2838:229–240.

29. Jakulin A, Bratko I, Smrke D, Demsar J, Zupan B: Attribute interactions in medical data analysis. Lect Notes Artif Intell 2003, 2780:229–238.
30. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B: Microarray data mining with visual programming. Bioinformatics 2005, 21:396–398.
31. Murphy MM, Vilella E, Ceruelo S, Figuera L, Sanchez M, Camps J, Cuco G, Ferre´ N, Labad A, Tasevska N, Arija V, Joven J, Fernandez-Ballart J: The MTHFR C677T, APOE, and PON55 gene polymorphisms show relevant interactions with cardiovascular risk factors. Clin Chem 2002, 48:372–375.
32. Ordovas JM: HDL genetics: candidate genes, genome wide scans and gene-environment interactions. Cardiovasc Drugs Ther 2002, 16:273–281.
33. Liao D, Tan H, Hui R, Li Z, Jiang X, Gaubatz J, Yang F, Durante W, Chan L, Schafer AI, Pownall HJ, Yang X, Wang H: Hyperhomocysteinemia decreases circulating high-density lipoprotein by inhibiting apolipoprotein A-I protein synthesis and enhancing HDL cholesterol clearance. Circ Res 2006, 99:598.
34. Enot DP, Beckmann M, Overy D, Draper J: Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of exploratory signals. PNAS 2006, 103:14865–14870.
35. Ritchie MD, Hahn LW, Moore JH: Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003, 24:150–157.
36. Province MA, Shannon WD, Rao DC: Classification methods for confronting heterogeneity. Adv Genet 2001, 42:273–286.
37. McKinney BA, Reif DM, Ritchie MD, Moore JH: Machine learning for detecting gene-gene interactions. Appl Bioinf 2006, 5: 77–88.
38. Moore JH, Williams SM: Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays 2005, 27:637–646.

Chapter 4

Sex-specific leptin-independent effects of CNTF, IL6 and UCP2 polymorphisms on weight gain.

A Geert Heidema[1,2,3]
Ping Wang[1]
Caroline TM van Rossum[2]
Edith JM Feskens[3]
Jolanda MA Boer[2]
Freek G Bouwman[1]
Pieter van 't Veer[3]
Edwin CM Mariman[1]

[1] Department of Human Biology, Maastricht University, Maastricht, The Netherlands;
[2] Centre for Nutrition and Health, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands;
[3] Division of Human Nutrition, Wageningen University and Research Centre, Wageningen, The Netherlands.

Submitted.

Abstract

The human proteins ciliary neurotrophic factor (CNTF) and interleukin-6 (IL6) and their receptors share structural homology with leptin and its receptor. Experiments have shown that CNTF and IL6, like leptin, can influence body weight in humans and animals. In a Dutch general population (n=545) we investigated effects of ciliary neurotrophic factor (CNTF null G/A, rs1800169), interleukin-6 (IL6 -174 G/C, rs1800795) and uncoupling protein-2 (UCP2 A55V, rs660339 and UCP2 del/ins) polymorphisms on weight gain using interaction graphs and logistic regression analysis. The average follow-up period was 6.9 years. Individuals who gained weight (n=264) were compared with individuals who remained stable in weight (n=281).

In women the CNTF polymorphism (odds ratio (OR)=2.15, 95%CI: 1.27-3.64, p=0.004) and in men the IL6 polymorphism by itself (OR=2.26, 95%CI: 1.08-4.75 , p=0.03) or in combination with the CNTF polymorphism, were associated with weight gain. Furthermore, CNTF and IL6 polymorphisms in interaction with UCP2 polymorphisms had similar strong effects on weight gain in women and men, respectively. All observed effects were independent of leptin. These results are incorporated in a biological model for weight regulation with upstream effects of CNTF and IL6, and downstream effects of UCP2.

The results of this study suggest a novel mechanism for weight regulation that is independent of leptin and active in both women and men, but strongly influenced by sex.

Introduction

Leptin is a well-studied hormone involved in energy metabolism. Mainly derived from adipocytes, leptin is known to be strongly related to body mass index (BMI) [1] and weight gain [2]. Besides in peripheral tissue [3, 4], leptin binds to its receptor in hypothalamic nuclei, thereby activating intracellular signaling pathways. Other proteins exist of which their receptors share structural homology with the leptin receptor. For example, the receptors of ciliary neurotrophic factor (CNTF) and interleukin-6 (IL6) have signal transduction elements similar to that present in the leptin receptor [5]. Both CNTF and IL6 receptors activate signaling pathways similar to those activated by the leptin receptor [6, 7]. Moreover, the VAST-program (http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml) shows that the human proteins leptin, CNTF and IL6 have high similarities in 3D-structure. CNTF and IL6 are therefore likely candidates to be involved in the regulation of energy homeostasis.

Administration of exogenous CNTF has shown to significantly reduce weight in animal models [7, 8]. After treatment with CNTF, obese mice that are deficient of functional leptin (ob/ob) or have a mutated leptin receptor (db/db) showed a reduction in food intake and body weight [7]. Moreover, CNTF is able to activate in hypothalamic nuclei leptin-like signaling pathways in diet-induced obese mouse-models that are not responsive to leptin [8]. In humans, treatment with recombinant variant CNTF resulted in more weight loss than placebo [9]. Thus, CNTF causes weight reduction in animal models and humans by leptin-like, but possibly leptin-independent pathway.

Of the cytokines, IL6 has many characteristics in common with leptin as an adiposity signal [10]. Besides in adipose tissue, IL6 is also expressed in hypothalamic nuclei involved in the regulation of energy homeostasis [11]. In the brain, IL6 decreases food intake [12] and enhances energy expenditure. Loss of circulating IL6 in IL6 null mice was associated with mature-onset obesity and low energy expenditure [13]. Central injection of IL6 in these animals led to increased energy expenditure. In humans, IL6 levels in the central nervous system (CNS) were found to be negatively correlated with fat mass [14].

In the CNTF gene, a G-to-A substitution at position -6 from the second exon was described by Takahashi et al. [15]. This splice mutation leads to a frame shift resulting in the total absence of functional CNTF protein in A/A homozygotes and approximately halved expression in heterozygotes [15]. In males homozygous for this CNTF null mutation, a higher body weight and BMI was observed [16]. As CNTF administration leads to weight loss, lack of CNTF due to the presence of the polymorphism may therefore result in decreased activation of anorectic pathways and consequently in an increase in body weight.

A polymorphism in the promoter region of the IL6 gene is the G-to-C substitution at position -174 [17]. The C-allele of this polymorphism was found to be associated with lower IL6 expression and lower plasma IL6 levels [17-19]. Subjects with the C/C genotype were shown to have lower energy expenditure [18]. In men, the IL6 -174 G/C polymorphism was also observed to be related with BMI [20], although this relationship is not consistently found [21]. Regarding weight gain, preliminary results from the present study indicated possible interaction between IL6 -174 G/C and uncoupling protein-2 (UCP2) polymorphisms. UCP2 is an inner mitochondrial membrane protein widely

expressed, including in white adipose tissue and skeletal muscle [22], and is likely involved in energy expenditure [23]. Although UCP2 polymorphisms were not found to be associated individually with weight gain [24], these polymorphisms may be involved in weight gain in interaction with CNTF null G/A and IL6 -174 G/C. Therefore, we investigated in this study the effect of CNTF null G/A, IL6 -174 G/C and their interactions with UCP2 polymorphisms on weight gain in a general population.

Methods

Study population

Two large population-based Dutch cohorts (Maastricht and Doetinchem) have been followed up between 1987 and 1998 [25]. For the present study, subjects were selected from these cohorts to obtain a group of subjects with stable weight and a group of subjects that gained weight during the follow-up period. Criteria to include and exclude subjects for participation in this study are shown in Box 1. More detailed information is provided by Van Rossum et al. [24].

Subjects with stable weight were frequency matched with weight gainers for age, cohort and smoking status. Subjects with missing observations for baseline leptin measurements (n=34) and subjects with missing observations for the CNTF polymorphism (n=3) were excluded. Totally 545 subjects were available for the analyses, 281 individuals with stable weight and 264 individuals that gained weight. Non-fasting blood samples were taken at the Municipal Health Services of Maastricht and Doetinchem. Both whole blood samples, in EDTA-coated vacutainer tubes, and EDTA-plasma samples were stored. Informed consent for using the stored blood samples for research purposes was given by all subjects.

Box 1

Criteria to include subjects:

- Subjects with age between 20 and 40 years
- Subjects that gained weight, the top decile of distribution of average weight gain per year
    - equal to, or more than 1.3 kg/year in Maastricht
    - equal to, or more than 1.4 kg/year in Doetinchem
- Subjects from both cohorts with stable weight, between -0.3 and 0.3 kg/year

Criteria to exclude subjects:

- Subjects using a energy or fat restricted diet
- Subjects using more than five glasses of alcoholic beverages per day
- Subjects with a chronic disease or serious illness
- Subjects who changed smoking habits recently
- Women who were pregnant at baseline or follow up
- Subjects with a follow-up of less than four years

Measurements

Anthropometric measurements

Weight at baseline and weight at the second measurement were measured at the Municipal Health Centre, with the exception of the second measurement for participants from Maastricht, which was self-reported (n=234). Subjects were measured at the Municipal Health Centre wearing light indoor clothing without shoes. For self-reported measurements it was assumed that subjects wore less clothing. Therefore, 1.5 kg was added to the weight of self-reported measurements. Weight at baseline was divided by height square (in kg/m^2) to obtain the body mass index (BMI) at baseline. The difference in weight between the second and first measurement was calculated and divided by the follow-up time (in years) to take differences in the follow-up period into account. This average annual weight gain was used as outcome measure.

Laboratory analyses

Leptin

Leptin concentrations were measured in duplicate in the non-fasting plasma samples by radioimmunoassay (HL-81K kit, Linco Research Inc., St. Charles, MO).

Genotyping of CNTF G/A null mutation, IL6 -174 G/C, UCP2 A55V and UCP2 del/ins

Genomic DNA was extracted from frozen buffy coats [26]. Genotyping of the G>A null mutation at position -6 before the second exon of the CNTF gene (rs1800169), was performed by using commercially available TaqMan SNP Genotyping assay (Applied Biosystems, Nieuwerkerk a/d IJsel, The Netherlands). The procedure was performed according the manufacturer's protocol and measured on an Applied Biosystems 7900 HT Fast Real-Time PCR System. Allelic calls were determined semi-automatically with the aid of the allelic discrimination software of Applied Biosystems.

UCP2 A55V (rs660339) has been determined by restriction fragment-length polymorphism analysis. The PCR products (UCP2 A55V: forward primer GGGCCAGTGCGACCTACAG and reverse primer ATGCGGACAGAGGCAAAGC) were digested with *Ecl* HK I, electrophoresed on a 3% agarose gel with ethidium bromide and visualized by UV transillumination.

IL6 -174 G/C (rs1800795) and UCP2 del/ins were determined by PCR-Pyrosequencing. The following primers were used for IL6: forward primer GCCTCAATGACGACCTAAGC, reverse biotin labeled primer TCATGGGAAAATCCCACATT and pyrosequence primer CCCCTAGTTGTGTCTTGC.

Primers used for UCP2 del/ins were: forward primer CAGTGAGGGAAGTGGGAGG, reverse biotin labeled primer GGGGCAGGACGAAGATTC. The biotinylated PCR products were immobilized on streptavidin-coated Sepharose beads (Amersham Biosciences). A sodium hydroxide solution was added to generate single-stranded DNA and the samples were washed and dissolved in annealing buffer. After adding the sequence primer, annealing was performed at 80°C for 3 min and the sequence reaction was performed automatically with a PSQ96 system by using a SNP reagent kit (Pyrosequencing AB, Sweden).

Statistical analysis

Hardy-Weinberg equilibrium for the different SNPs was calculated in the weight stable group, using the chi-square test. The CNTF, IL6 and UCP2 del/ins polymorphisms were in Hardy-Weinberg equilibrium (HWE). For UCP2 A55V deviation from HWE was observed ($p=0.01$). This deviation is not very strong and therefore we included UCP2 A55V in the analyses. Linkage disequilibrium (LD) between the UCP2 polymorphisms was present ($r=0.63$, $p<0.0001$).

All further statistical analyses were performed for men and women separately and adjusted for matching criteria, i.e. age, cohort and smoking status. The relationship between each of the CNTF and IL6 polymorphisms and weight gain was analyzed by the chi-square test. Logistic regression analyses were performed to study the effect of these polymorphisms and their interaction effect with leptin on weight gain, and to obtain risk estimates for the multi-locus genotypes for combinations of CNTF, IL6 and UCP2 polymorphisms that were shown to have an important effect on weight gain by the interaction graph (see below). Weight at baseline was also included in the logistic regression model, as a significant difference between the stable weight and weight gainers group was observed. To perform the analyses SAS software version 9.1 (SAS institute Inc., Cary, NC, USA) was used.

Interaction entropy graph

To obtain a measure of importance for the different SNP-SNP interactions we obtained an interaction entropy graph [27-29], for men and women separately. This graphical model is useful to compare the independent main effects of polymorphisms with the pair-wise interaction effect between polymorphisms. Positive or negative entropy values for pair-wise SNP interactions indicate that besides the two individual SNP effects information is explained by their interaction. Positive entropy values for interactions indicate a synergistic interaction whereas negative entropy values indicate the presence of redundant information. To create the entropy-based interaction graph the Orange machine learning software was used, which is open-source and freely available [30] at http://www.ailab.si/orange.

Results

Study population

The mean age at baseline for the individuals in this study population was 29.3 years, with a range from 20.2 to 39.8 years (table 1). BMI at baseline was on average 23.0 kg/m^2, ranging from 16.3 to 37.8 kg/m^2.

Table 1: Characteristics of the study population.

Population characteristics	Stable weight		Weight gain	
	Men	Women	Men	Women
Number of subjects	134	147	127	137
Baseline				
Age (years)	28.9 ± 5.6	29.9 ± 6.0	28.1 ± 5.9	30.0 ± 5.9
Weight (kg)	76.6 ± 10.0	63.5 ± 8.1	79.4 ± 10.9	66.3 ± 11.5
Height (cm)	1.80 ± 0.07	1.67 ± 0.06	1.82 ± 0.07	1.67 ± 0.07
BMI (kg/m^2)	23.3 ± 2.8	22.3 ± 2.8	23.6 ± 3.0	23.1 ± 3.9
Leptin (μg/L)	3.47 ± 2.19	10.6 ± 6.07	4.02 ± 2.56	13.5 ± 10.8
Smoking status (%)	35.1	33.3	36.2	35.0
Follow-up				
Follow-up time (years)	7.2 ± 1.9	6.8 ± 1.6	6.9 ± 1.7	6.7 ± 1.5
Weight (kg)	75.5 ± 10.0	62.5 ± 8.0	90.5 ± 11.5	77.6 ± 12.5
Weight gain (kg/year)	0.06 ± 0.17	0.08 ± 0.16	1.86 ± 0.54	1.95 ± 0.60
BMI (kg/m^2)	23.4 ± 2.8	22.5 ± 2.8	27.4 ± 3.3	27.7 ± 4.3

Results are presented as mean ± SD, except for smoking status which is given in percentages.

Effect of CNTF and IL6 polymorphisms on weight gain

The interaction entropy graphs (figure 1A and 1B) show the individual effects of CNTF, IL6 and the two UCP2 polymorphisms and the pair-wise interactions between these polymorphisms on weight gain for men and women, respectively. For men, the amount of entropy explained by IL6 -174 G/C is 1.32% and a smaller effect was found for CNTF G/A (1.01%). CNTF G/A in combination with IL6 -174 G/C explained 2.36% in addition to the individual effects, suggesting interaction between these two polymorphisms. For women, the CNTF polymorphism explained 2.11% of the entropy, whereas an effect of the IL6 -174 G/C is absent (0.58%). Also, no interaction effect between CNTF G/A and IL6 -174 G/C is present (0.41%). These results of the interaction graphs are in accordance with the results of traditional methods. In table 2 the allele and genotype frequencies for CNTF G/A and IL6 -174 G/C are shown. An association between IL6 -174 G/C and weight gain was observed in men, a tendency for weight gain when taken dichotomously ($p=0.09$) and a significant effect when taken as a continuous trait ($p=0.03$). In women, a clear significant association of CNTF G/A with weight gain was found, both when taken dichotomously ($p=0.004$) and continuously ($p=0.0008$).

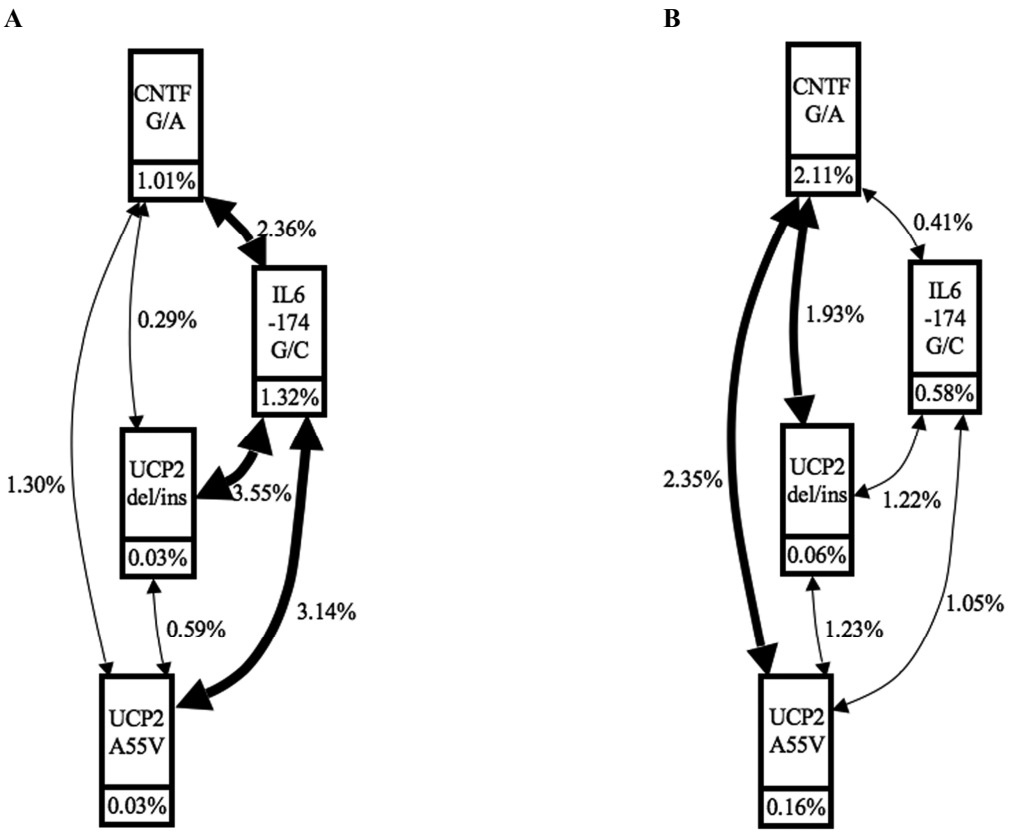


Figure 1A and 1B: Entropy based interaction graph for men and women, respectively. The percentages in the boxes represent the amount of entropy of weight gain explained by the individual SNP-effects, whereas the percentages at the arrows represent the amount of entropy of weight gain explained by pair-wise interaction effects, on top of the two individual SNP effects. Positive entropy values for pair-wise SNP interactions indicate a synergistic interaction whereas negative entropy values indicate the presence of redundant information.

Similar results were obtained using logistic regression analyses (adjustment for age, cohort, smoking status and weight at baseline did not alter the results). Men with the IL6 -174 C/C genotype had an increased risk for weight gain (odds ratio (OR)=2.26, 95%CI: 1.08-4.75, p=0.03) compared to men with the G/G genotype. Including interaction between CNTF and IL6 polymorphisms in the model, significant multi-locus effects were observed in men. The nine multi-locus genotypes are depicted in figure 2.

Table 2: Relationship between CNTF G/A and IL6 -174 G/C polymorphisms with weight gain dichotomous and weight gain continuous.

		Men		Weight gain/year		Women		Weight gain/year	
SNP	genotype	stable weight (n=134)	weight gainers (n=127)	χ^2 p-value	continuous p-value	stable weight (n=147)	weight gainers (n=137)	χ^2 p-value	continuous p-value
CNTF G/A	G/G	100 (75%)	92 (72%)	0.17	0.40	116 (79%)	87 (64%)	0.004	0.0008
	G/A	27 (20%)	33 (26%)			31 (21%)	50 (36%)		
	A/A	7 (5%)	2 (2%)			-	-		
	allele								
	G	0.85	0.85	0.81		0.89	0.82	0.009	
	A	0.15	0.15			0.11	0.18		
	genotype								
IL6 -174 G/C	G/G	60 (45%)	46 (36%)	0.09	0.03	57 (39%)	56 (41%)	0.32	0.33
	G/C	59 (44%)	55 (43%)			74 (50%)	59 (43%)		
	C/C	15 (11%)	26 (20%)			16 (11%)	22 (16%)		
	allele								
	G	0.67	0.58	0.04		0.64	0.62	0.70	
	C	0.33	0.42			0.36	0.38		

Conclusions for multi-locus genotypes including the CNTF A/A genotype and for the multi-locus genotype CNTF G/A with IL6 -174 C/C can not be drawn as the number of subjects for these multi-locus genotypes are too small. The results show that having two mutant alleles in IL6 (OR=2.74, 95%CI: 1.14-6.59, p=0.02) or a mutant allele for both CNTF and IL6 increases the risk of weight gain (OR=2.98, 95%CI: 1.10-8.07, p=0.03). For women, an increased risk for weight gain was observed for the CNTF G/A genotype (OR=2.15, 95%CI: 1.27-3.64, p=0.004). No significant association with weight gain was observed for the combination of CNTF and IL6 polymorphisms.

Including baseline leptin level in the model did not alter the results for both men and women. Furthermore, interaction effects of CNTF G/A with leptin (p=0.65 and p=0.55 for men and women, respectively), or IL6 -174 G/C with leptin (p=0.97 and p=0.28 for men and women, respectively) on weight gain were found to be non-significant.

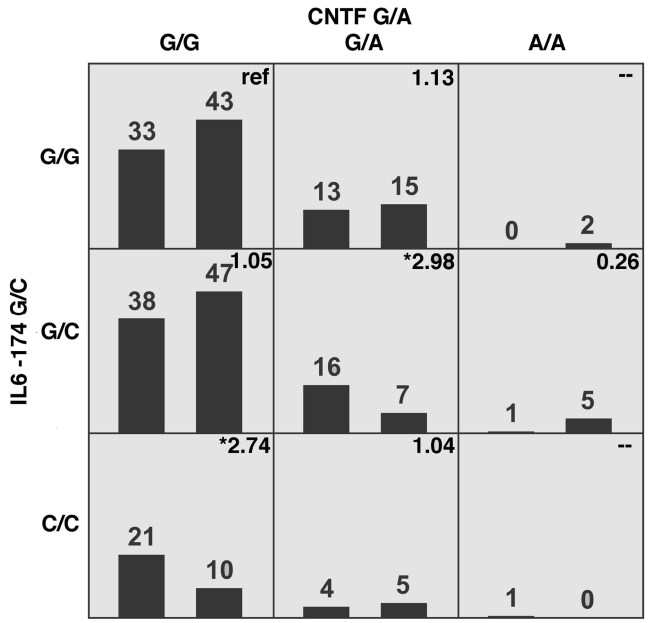


Figure 2: Combination of CNTF G/A with IL6 -174 G/C in men.
* Significant at the 0.05 level
** Significant at the 0.01 level
ref: reference group
In each cell, the left bar indicates the number of weight gainers and the right bar the number of stable weight individuals. The odds ratio estimate is shown at the upper right corner.

Effect of CNTF and IL6 polymorphisms in interaction with UCP2 polymorphisms on weight gain

The interaction graphs (figure 1A and 1B) confirm previous results [24] that main effects of both UCP2 polymorphisms on weight gain were absent in both men and women. However, in men, the amount of entropy explained by the interactions between IL6 -174

G/C and UCP2 A55V and between IL6 -174 G/C and UCP2 del/ins was high (3.14% and 3.55%, respectively), indicating interaction effects between these polymorphisms on weight gain. In women, interaction effects between CNTF G/A and the UCP2 polymorphisms were present (1.93% and 2.35%).

For men, the combination of IL6 -174 G/C with the UCP2 polymorphisms are depicted in figure 3A and 3B. A strongly increased risk is observed for individuals who have the IL6 -174 C/C genotype in combination with the wild-type homozygous or heterozygous genotype for UCP2 A55V (OR=4.57, 95%CI: 1.54-13.6, p=0.006 and OR=5.33, 95%CI: 1.44-19.8, p=0.01, respectively). A significantly increased risk of weight gain is also observed for the multi-locus genotypes of IL6 -174 G/C with UCP2 55 A/A (OR=2.49, 95%CI: 1.10-5.65, p=0.03) and IL6 -174 G/G with UCP2 55 V/V (OR=3.35, 95%CI: 1.08-10.4, p=0.04). The frequency bars show a similar pattern for the combination of IL6 polymorphism with the UCP2 del/ins polymorphism (figure 3B), although only the multi-locus genotype of IL6 -174 C/C with UCP2 del/ins was found to be significant at the 0.05 level (OR=8.25, 95%CI: 1.69-40.3, p=0.009).

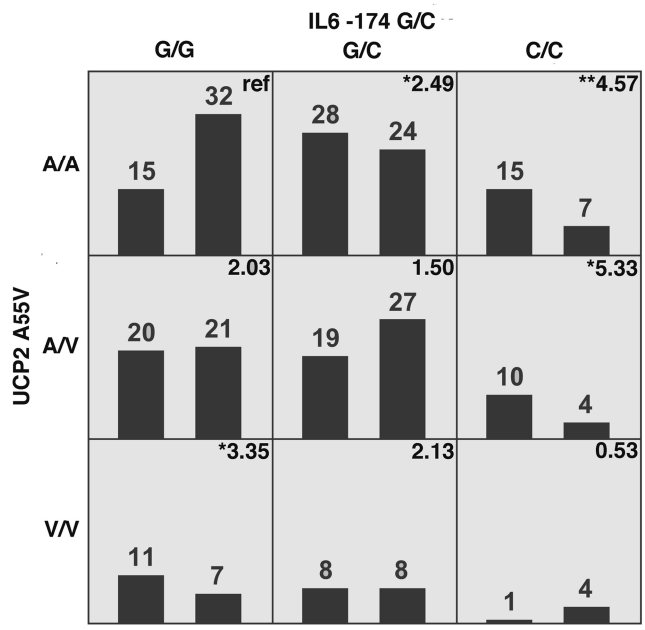


Figure 3A: Combination of IL6 -174 G/C with UCP2 A55V in men.
* Significant at the 0.05 level
** Significant at the 0.01 level
ref: reference group
In each cell, the left bar indicates the number of weight gainers and the right bar the number of stable weight individuals. The odds ratio estimate is shown at the upper right corner.

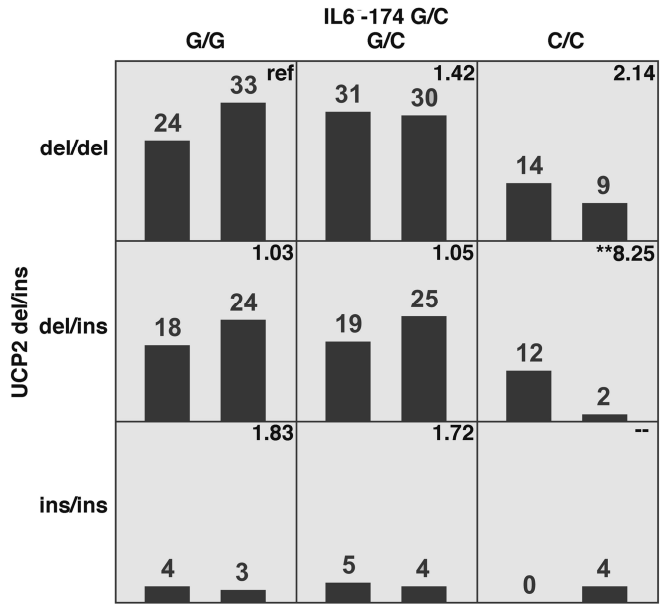


Figure 3B: Combination of IL6 -174 G/C with UCP2 del/ins in men.
* Significant at the 0.05 level
** Significant at the 0.01 level
ref: reference group
In each cell, the left bar indicates the number of weight gainers and the right bar the number of stable weight individuals. The odds ratio estimate is shown at the upper right corner.

For women, the combinations of CNTF G/A with the UCP2 polymorphisms are shown in figure 4A and 4B. From these figures it can be seen that a significantly increased risk is present for the group with one mutant allele for the CNTF polymorphism in combination with the wild-type or heterozygous genotype for either of the UCP2 polymorphisms. This risk pattern corresponds with the risk pattern observed in men, only, instead of the IL6 -174 G/C polymorphism, the CNTF polymorphism plays a role in women. In women with one mutant allele for CNTF, a significant increased risk is observed in combination with UCP2 55 A/A (OR=6.98, 95%CI: 2.54-19.2, p=0.0002) and UCP2 55 A/V (OR=2.28, 95%CI: 1.02-5.08, p=0.04). This increased risk was also observed in women with one mutant allele for CNTF in combination with UCP2 del/del (OR=4.18, 95%CI: 1.92-9.13, p=0.0003) and UCP2 del/ins (OR=2.30, 95%CI: 1.03-5.12, p=0.04).

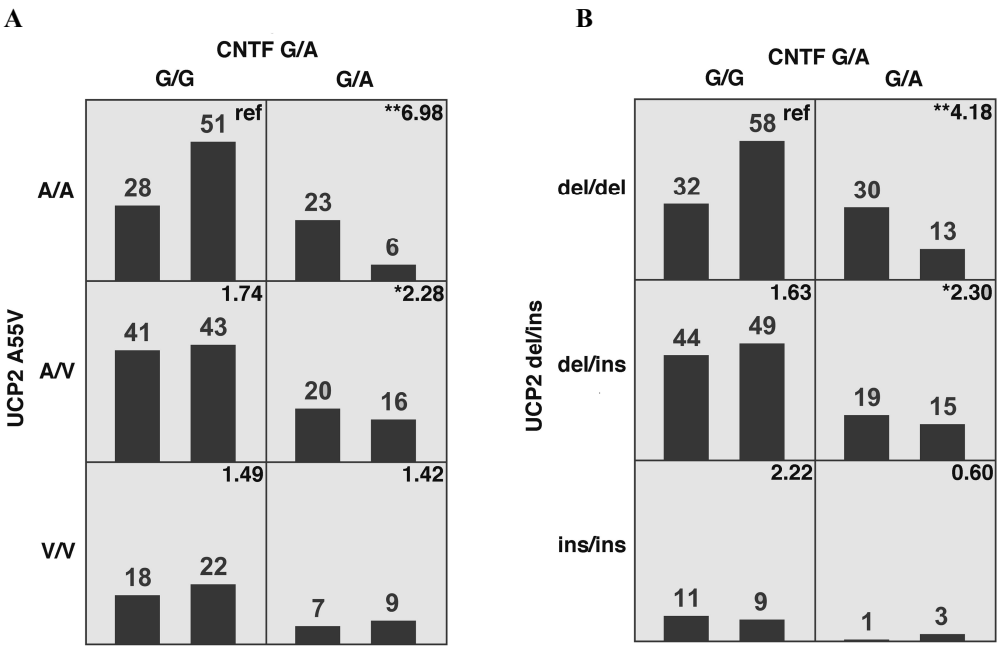


Figure 4A: Combination of CNTF G/A with UCP2 A55V in women.
Figure 4B: Combination of CNTF G/A with UCP2 del/ins in women.
* Significant at the 0.05 level
** Significant at the 0.01 level
ref: reference group
In each cell, the left bar indicates the number of weight gainers and the right bar the number of stable weight individuals. The odds ratio estimate is shown at the upper right corner.

Discussion

We report in a Dutch general population the sex-specific effects of CNTF G/A, IL6 -174 G/C and their pair-wise interactions with UCP2 polymorphisms on weight gain. CNTF and IL6 individually are associated with weight gain in this cohort in women and men, respectively. As shown by Kubaszek et al. [18], subjects with the C/C genotype of IL6 -174 polymorphism had significantly lower energy expenditure and basal metabolic rate compared to carriers of the G allele, which corresponds to the results obtained for men in this study. Although in their study this effect was measured in men and women together, the number of men in their study outnumbered the number of women. In our study no main effect for CNTF was present in men, but the combination of CNTF and IL6 increased the risk for weight gain. The number of observations present for some multi-locus genotypes was not sufficiently large to draw strong conclusions. However, together these results indicate that in men IL6 is more potent than CNTF in weight regulation and that two mutant alleles in IL6 or at least one mutant allele in both CNTF and IL6 polymorphisms increase the risk of weight gain.

No interaction effects between CNTF and IL6 were found in women, while a main effect was present for the CNTF G/A genotype. The IL6 polymorphism was not associated with weight gain in women. This can be due to the high estrogen level present in premenopausal women that blunts the effect of IL6 [31, 32]. In our study population the presence of one mutant allele in the CNTF polymorphism, most likely lowering the concentration of CNTF, in women was already a sufficient condition to increase the risk for weight gain.

From the results of both men and women it is tentative to speculate that in general a strong decrease in IL6 protein concentrations or a mild decrease in both CNTF and IL6 proteins is sufficient to increase the risk for weight gain. This hypothesis suggests a biological model in which IL6 and CNTF cross-activate each other's receptors and/or activate parallel energy expenditure pathways (figure 5). As the human proteins CNTF and IL6 have high similarities in 3D-structure, as can been seen from the VAST-program (http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml), and their receptors show similar signaling capabilities [6], cross-activation seems reasonable. The VAST-program also shows structural homology between leptin and CNTF or IL6. However, associations between CNTF and IL6 polymorphisms and weight gain did not change after adjusting for leptin at baseline in the model. Furthermore, interaction effects between these polymorphisms and leptin were non-significant. Therefore, CNTF and IL6 presumably influence weight regulation independently from leptin.

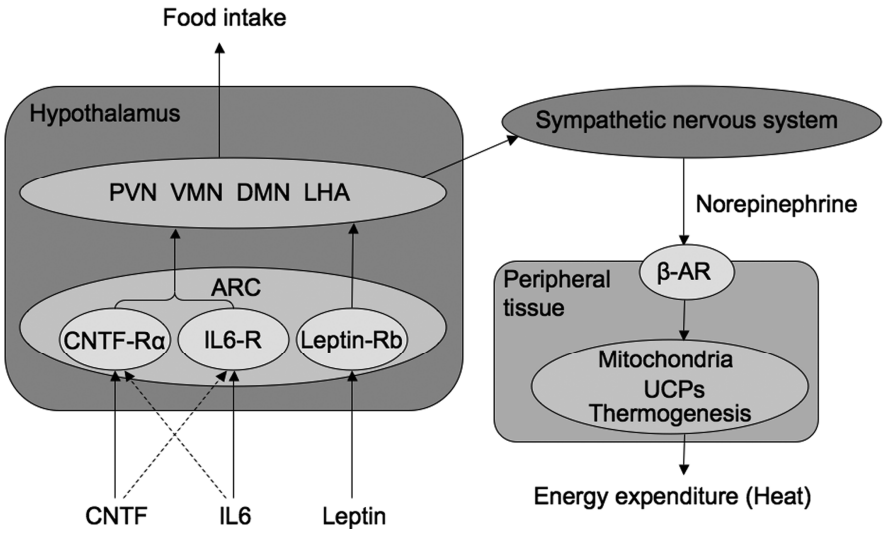


Figure 5: Biological model including the effect of CNTF, IL6, leptin and UCP2 on energy expenditure. For simplicity, direct effects of IL6 and leptin on peripheral tissue are not included. Dashed lines indicate that CNTF and IL6 may cross-activate each other's receptors. The interaction between CNTF and IL6 is present in men, whereas IL6 is blunted in women due to high estrogen levels. ARC: arcuate nucleus. PVN: paraventricular nucleus. VMN: ventricular medial nucleus. DMN: dorsomedial nucleus. LHA: lateral hypothalamic area. β-AR: beta adrenergic receptor.

In this study, the UCP2 polymorphisms were not individually associated with weight gain. This is in line with other studies in which UCP2 polymorphisms were not associated with obesity-related phenotypes [24, 33, 34]. However, taking interactions between polymorphisms into account, UCP2 polymorphisms were found to have an effect in combination with CNTF and IL6 polymorphisms in women and men, respectively. This finding shows that relying only on univariate analyses is not sufficient for the detection of important SNPs. We found that the multi-locus genotypes of IL6 -174 C/C with UCP2 55 A/A and A/V in men and CNTF G/A with UCP2 55 A/A and A/V in women show a significantly increased risk for weight gain. In men strong interactions of IL6 with UCP2 polymorphisms are present, but not between CNTF with UCP2. This is also an indication that IL6 may be more dominant than CNTF in weight regulation. As IL6 in women may be blunted, only interactions between CNTF and UCP2 are present in women.

The interactions of CNTF and IL6 with UCP2 have an effect on weight gain most likely via influencing energy expenditure. CNTF, IL6, and leptin have an effect on hypothalamic nuclei, which in turn alter the activity of the sympathetic nervous system (SNS). By the secretion of norepinephrine, the SNS controls the UCP2 expression and thereby thermogenesis in peripheral tissue [35]. Studies in rats demonstrate that leptin increases sympathetic activation of UCP1 gene expression in brown adipose tissue [36]. IL6 administration to CNS also increases sympathetic induction of UCP1 in brown adipose tissue, reducing body weight [37]. Whereas UCP1 is primarily expressed in brown adipose tissue, UCP2 is widely expressed including in white adipose tissue and skeletal muscle [38]. As UCP2 is structurally homologous to UCP1 [38], similar effects can be expected for UCP2. Indeed, central infusion of leptin was shown to increase the expression of UCP2 in white adipose tissue [39]. Similar effects for CNTF and IL6 may be expected.

Some issues regarding this study should be considered. First, for IL6, CNTF and UCP2, results were obtained at the SNP level and not the protein level. Biological interpretations at the protein levels should therefore be considered as indicative. However, CNTF [15] and IL6 [17-19] polymorphisms were found to influence their corresponding protein levels. Therefore it is likely that their results correspond to the protein level. Secondly, in this study UCP2 A55V was not in HWE, therefore conclusions for this polymorphism should be considered cautiously. However, similar results were obtained for the UCP2 del/ins polymorphism, which was in HWE. These UCP2 polymorphisms were in LD. Therefore it is likely that a true association between UCP2 A55V, in interaction with either CNTF or IL6 polymorphisms, and weight gain exists. Also, weight at the second measurement was self reported for participants from Maastricht. However, the distribution of weight gainers and weight stable group over the CNTF and IL6 genotypes was similar between the two cohorts (Doetinchem and Maastricht), in both men and women, suggesting that the self-reported measurements did not influence the associations found in this study. Furthermore, self-reported measurements are invariably under reported [40], resulting in an attenuation of the observed associations. Therefore, true associations would be even stronger than the associations found in our study.

In conclusion, CNTF and IL6 polymorphisms are sex-specifically related to weight gain, independently of circulating leptin level. CNTF and IL6 polymorphisms in interaction with UCP2 polymorphisms also have strong effects on weight gain, in women and men respectively. These results suggest that CNTF and IL6, via leptin-independent pathways,

together with UCP2 influence energy expenditure and thereby weight gain in a sex-specific manner.

Acknowledgements

This project has been carried out in the framework of the Centre for Human Nutrigenomics. The Cardiovascular Disease Risk Factor Monitoring Study (PEILSTATION) and the Monitoring Project on Risk Factors for Chronic Diseases (MORGEN-project) were financially supported by the Ministry of Health, Welfare and Sport of The Netherlands and the National Institute for Public Health and the Environment. We thank the epidemiologists and field workers of the Municipal Health Services in Doetinchem and Maastricht for their contribution to the data collection, and G.L. Obermann-de Boer for coordinating the PEILSTATION-project. The project steering committee of MORGEN project consisted of Dr. H.B. Bueno de Mesquita, Dr. H.A. Smit, Dr. W.M.M. Verschuren and Prof. Dr. Ir. J.C. Seidell. Logistic management was provided by A. Jansen and J. Steenbrink and data management by A. Blokstra, P. Steinberger and A. van Kessel. We thank Dr. B. Hoebee and Prof. Dr. Ir. J.C. Seidell for their contribution to the conception, design and realization of the original study on weight gain.

References

1. Considine RV, Sinha MK, Heiman ML, Kriauciunas A, Stephens TW, Nyce MR, Ohannesian JP, Marco CC, McKee LJ, Bauer TL et al: Serum immunoreactive-leptin concentrations in normal-weight and obese humans. N Engl J Med 1996, 334(5):292-295.
2. van Rossum CT, Hoebee B, van Baak MA, Mars M, Saris WH, Seidell JC: Genetic variation in the leptin receptor gene, leptin, and weight gain in young Dutch adults. Obes Res 2003, 11(3):377-386.
3. Friedman JM, Halaas JL: Leptin and the regulation of body weight in mammals. Nature 1998, 395(6704):763-770.
4. Minokoshi Y, Kim YB, Peroni OD, Fryer LG, Muller C, Carling D, Kahn BB: Leptin stimulates fatty-acid oxidation by activating AMP-activated protein kinase. Nature 2002, 415(6869):339-343.
5. Plata-Salaman CR: Cytokine-induced anorexia. Behavioral, cellular, and molecular mechanisms. Ann N Y Acad Sci 1998, 856:160-170.
6. Baumann H, Morella KK, White DW, Dembski M, Bailon PS, Kim H, Lai CF, Tartaglia LA: The full-length leptin receptor has signaling capabilities of interleukin 6-type cytokine receptors. Proc Natl Acad Sci U S A 1996, 93(16):8374-8378.
7. Gloaguen I, Costa P, Demartis A, Lazzaro D, Di Marco A, Graziani R, Paonessa G, Chen F, Rosenblum CI, Van der Ploeg LH et al: Ciliary neurotrophic factor corrects obesity and diabetes associated with leptin deficiency and resistance. Proc Natl Acad Sci U S A 1997, 94(12):6456-6461.
8. Lambert PD, Anderson KD, Sleeman MW, Wong V, Tan J, Hijarunguru A, Corcoran TL, Murray JD, Thabet KE, Yancopoulos GD et al: Ciliary neurotrophic factor activates leptin-like pathways and reduces body fat, without cachexia or rebound weight gain, even in leptin-resistant obesity. Proc Natl Acad Sci U S A 2001, 98(8):4652-4657.

9. Ettinger MP, Littlejohn TW, Schwartz SL, Weiss SR, McIlwain HH, Heymsfield SB, Bray GA, Roberts WG, Heyman ER, Stambler N et al: Recombinant variant of ciliary neurotrophic factor for weight loss in obese adults: a randomized, dose-ranging study. Jama 2003, 289(14):1826-1832.
10. Cancello R, Tounian A, Poitou C, Clement K: Adiposity signals, genetic and body weight regulation in humans. Diabetes Metab 2004, 30(3):215-227.
11. Schobitz B, de Kloet ER, Sutanto W, Holsboer F: Cellular localization of interleukin 6 mRNA and interleukin 6 receptor mRNA in rat brain. Eur J Neurosci 1993, 5(11):1426-1435.
12. Wallenius K, Wallenius V, Sunter D, Dickson SL, Jansson JO: Intracerebroventricular interleukin-6 treatment decreases body fat in rats. Biochem Biophys Res Commun 2002, 293(1):560-565.
13. Wallenius V, Wallenius K, Ahren B, Rudling M, Carlsten H, Dickson SL, Ohlsson C, Jansson JO: Interleukin-6-deficient mice develop mature-onset obesity. Nat Med 2002, 8(1):75-79.
14. Stenlof K, Wernstedt I, Fjallman T, Wallenius V, Wallenius K, Jansson JO: Interleukin-6 levels in the central nervous system are negatively correlated with fat mass in overweight/obese subjects. J Clin Endocrinol Metab 2003, 88(9):4379-4383.
15. Takahashi R, Yokoji H, Misawa H, Hayashi M, Hu J, Deguchi T: A null mutation in the human CNTF gene is not causally related to neurological diseases. Nat Genet 1994, 7(1):79-84.
16. O'Dell SD, Syddall HE, Sayer AA, Cooper C, Fall CH, Dennison EM, Phillips DI, Gaunt TR, Briggs PJ, Day IN: Null mutation in human ciliary neurotrophic factor gene confers higher body mass index in males. Eur J Hum Genet 2002, 10(11):749-752.
17. Fishman D, Faulds G, Jeffery R, Mohamed-Ali V, Yudkin JS, Humphries S, Woo P: The effect of novel polymorphisms in the interleukin-6 (IL-6) gene on IL-6 transcription and plasma IL-6 levels, and an association with systemic-onset juvenile chronic arthritis. J Clin Invest 1998, 102(7):1369-1376.
18. Kubaszek A, Pihlajamaki J, Punnonen K, Karhapaa P, Vauhkonen I, Laakso M: The C-174G promoter polymorphism of the IL-6 gene affects energy expenditure and insulin sensitivity. Diabetes 2003, 52(2):558-561.
19. Terry CF, Loukaci V, Green FR: Cooperative influence of genetic polymorphisms on interleukin 6 transcriptional regulation. J Biol Chem 2000, 275(24):18138-18144.
20. Berthier MT, Paradis AM, Tchernof A, Bergeron J, Prud'homme D, Despres JP, Vohl MC: The interleukin 6-174G/C polymorphism is associated with indices of obesity in men. J Hum Genet 2003, 48(1):14-19.
21. Lieb W, Pavlik R, Erdmann J, Mayer B, Holmer SR, Fischer M, Baessler A, Hengstenberg C, Loewel H, Doering A et al: No association of interleukin-6 gene polymorphism (-174 G/C) with myocardial infarction or traditional cardiovascular risk factors. Int J Cardiol 2004, 97(2):205-212.
22. Saleh MC, Wheeler MB, Chan CB: Uncoupling protein-2: evidence for its function as a metabolic regulator. Diabetologia 2002, 45(2):174-187.
23. Pecqueur C, Couplan E, Bouillaud F, Ricquier D: Genetic and physiological analysis of the role of uncoupling proteins in human energy homeostasis. J Mol Med 2001, 79(1):48-56.
24. van Rossum CT, Hoebee B, Seidell JC, Bouchard C, van Baak MA, de Groot CP, Chagnon M, de Graaf C, Saris WH: Genetic factors as predictors of weight gain in young adult Dutch men and women. Int J Obes Relat Metab Disord 2002, 26(4):517-528.
25. Verschuren WMM VLE, Blokstra A, Seidell JC, Smit HA, Bueno de Mesquita HB, Obermann-de Boer GL, Kromhout D.: Cardiovascular disease risk factors in The Netherlands. Neth J Cardiol 1993, 6:205-210.

26. Hoebee B, Rietveld E, Bont L, Oosten M, Hodemaekers HM, Nagelkerke NJ, Neijens HJ, Kimpen JL, Kimman TG: Association of severe respiratory syncytial virus bronchiolitis with interleukin-4 and interleukin-4 receptor alpha polymorphisms. J Infect Dis 2003, 187(1):2-11.
27. Jakulin A, Bratko I: Analyzing attribute dependencies. Lect Notes Artif Intell 2003, 2838:229-240.
28. Jakulin A, Bratko I, Smrke D, Demsar J, Zupan B: Attribute interactions in medical data analysis. Lect Notes Artif Intell 2003, 2780:229-238.
29. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 2006, 241(2):252-261.
30. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B: Microarray data mining with visual programming. Bioinformatics 2005, 21(3):396-398.
31. Masiukiewicz US, Mitnick M, Grey AB, Insogna KL: Estrogen modulates parathyroid hormone-induced interleukin-6 production in vivo and in vitro. Endocrinology 2000, 141(7):2526-2531.
32. Masiukiewicz US, Mitnick M, Gulanski BI, Insogna KL: Evidence that the IL-6/IL-6 soluble receptor cytokine system plays a role in the increased skeletal sensitivity to PTH in estrogen-deficient women. J Clin Endocrinol Metab 2002, 87(6):2892-2898.
33. Dalgaard LT, Sorensen TI, Andersen T, Hansen T, Pedersen O: An untranslated insertion variant in the uncoupling protein 2 gene is not related to body mass index and changes in body weight during a 26-year follow-up in Danish Caucasian men. Diabetologia 1999, 42(12):1413-1416.
34. Urhammer SA, Dalgaard LT, Sorensen TI, Moller AM, Andersen T, Tybjaerg-Hansen A, Hansen T, Clausen JO, Vestergaard H, Pedersen O: Mutational analysis of the coding region of the uncoupling protein 2 gene in obese NIDDM patients: impact of a common amino acid polymorphism on juvenile and maturity onset forms of obesity and insulin resistance. Diabetologia 1997, 40(10):1227-1230.
35. Bachman ES, Dhillon H, Zhang CY, Cinti S, Bianco AC, Kobilka BK, Lowell BB: betaAR signaling required for diet-induced thermogenesis and obesity resistance. Science 2002, 297(5582):843-845.
36. Scarpace PJ, Matheny M: Leptin induction of UCP1 gene expression is dependent on sympathetic innervation. Am J Physiol 1998, 275(2 Pt 1):E259-264.
37. Li G, Klein RL, Matheny M, King MA, Meyer EM, Scarpace PJ: Induction of uncoupling protein 1 by central interleukin-6 gene delivery is dependent on sympathetic innervation of brown adipose tissue and underlies one mechanism of body weight reduction in rats. Neuroscience 2002, 115(3):879-889.
38. Fleury C, Neverova M, Collins S, Raimbault S, Champigny O, Levi-Meyrueis C, Bouillaud F, Seldin MF, Surwit RS, Ricquier D et al: Uncoupling protein-2: a novel gene linked to obesity and hyperinsulinemia. Nat Genet 1997, 15(3):269-272.
39. Tajima D, Masaki T, Hidaka S, Kakuma T, Sakata T, Yoshimatsu H: Acute central infusion of leptin modulates fatty acid mobilization by affecting lipolysis and mRNA expression for uncoupling proteins. Exp Biol Med (Maywood) 2005, 230(3):200-206.
40. Sherry B, Jefferds ME, Grummer-Strawn LM: Accuracy of adolescent self-report of height and weight in assessing overweight status: a literature review. Arch Pediatr Adolesc Med 2007, 161(12):1154-1161.

Chapter 5

A framework to identify physiological responses in microarray-based gene expression studies: selection and interpretation of biologically relevant genes

A Geert Heidema[1,2,3] *
Wendy Rodenburg[4,5,6] *
Jolanda MA Boer[1]
Ingeborg MJ Bovee-Oudenhoven[4,6]
Edith JM Feskens[3]
Edwin CM Mariman[2]
Jaap Keijer[4,5]

[1] Centre for Nutrition and Health, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands;
[2] Department of Human Biology, Maastricht University, Maastricht, The Netherlands;
[3] Division of Human Nutrition, Wageningen University and Research Centre, Wageningen, The Netherlands;
[4] TI Food and Nutrition, Wageningen, The Netherlands;
[5] RIKILT-Institute of Food Safety, Wageningen, The Netherlands;
[6] NIZO Food Research, Ede, The Netherlands.

*These authors contributed equally to this paper.

Physiological Genomics 2008, 33(1):78-90.

Abstract

In whole genome microarray studies major gene expression changes are easily identified, but it is a challenge to capture small, but biologically important, changes. Pathway-based programs can capture small effects but may have the disadvantage of being restricted to functionally annotated genes. A structured approach toward the identification of major and small changes for interpretation of biological effects is needed. We present a structured approach, a framework, that addresses different considerations in 1) the identification of informative genes in microarray datasets and 2) the interpretation of their biological relevance. The steps of this framework include gene ranking, gene selection, gene grouping, and biological interpretation. Random forests (RF), which takes gene-gene interactions into account, is examined to rank and select genes. For human, mouse, and rat whole genome arrays, less than half of the probes on the array are annotated. Consequently, pathway analysis tools ignore half of the information present in the microarray dataset. The framework described takes all genes into account. RF is a useful tool to rank genes by taking interactions into account. Applying a permutation approach, we were able to define an objective threshold for gene selection. RF combined with self-organizing maps identified genes with coordinated but small gene expression responses that were not fully annotated but corresponded to the same biological process. The presented approach provides a flexible framework for biological interpretation of microarray datasets. It includes all genes in the dataset, takes gene-gene interactions into account, and provides an objective threshold for gene selection.

Introduction

Transcriptome analysis using whole genome microarrays is an elegant and widely used approach for identification of the molecular mechanisms underlying diet-induced cellular or physiological changes. Both major effects as well as a wide overview of more subtle changes can be obtained. While the major differences are important for classification and identification of individual response genes, the smaller changes are an integral part of the physiological response and are essential for the identification of the physiological processes that are affected by the challenge or intervention. This is especially true in nutrition, where dietary interventions result in modest, but biologically important gene expression changes [1-3]. In the medical field it is also increasingly recognized that the more subtle changes contribute importantly to the outcome [4-6].

To translate microarray data into functional physiological information, a set of genes with the maximum amount of information and a minimum of noise is needed. Although a large number of methods exist to select genes from microarray datasets, most methods aim to identify the smallest possible set of genes that still can discriminate, for example, to classify malignancies, predict therapeutic outcomes, or diagnose physiological responses [6, 7]. These methods may not always be appropriate to select a larger set of genes for biological interpretation that includes the smaller changes. These smaller changes are part of the response to medication or disease, which occurs through the interactions of multiple genes, via signaling pathways or other functional relationships. Small changes, variability among individuals, and the often small sample sizes on one hand and the large number of genes tested on the other make it difficult to distinguish true differences from noise [4, 8]. Careful planning and execution of microarray experiments nowadays offers technically high-quality data, with a minimum of noise. However, the combination of small gene expression changes and the needed selection of the largest informative set of genes demands sophisticated selection methods. A structured framework that incorporates the different considerations in the identification of informative genes and the interpretation of their biological relevance is needed. Here we describe the steps of such a framework and address the following considerations: gene ranking, gene selection, gene grouping, and biological interpretation.

Gene ranking

To identify genes of relevance within the total dataset, genes are ranked by a measure of importance. As such, fold change has often been used. However, fold change is not a reliable measure because it does not take variability in the data into account [9, 10]. Therefore, other measures that do take variability into consideration should be used. The most commonly used approach for gene selection in two-class microarray studies that takes variability into account is the conventional t-test, while ANOVA is used for multi-class studies. Genes are tested independently, and a p-value is assigned to each gene, which can be used to rank genes by their importance. However, by ranking genes by a univariate test-statistic such as the t-statistic, all genes in the dataset are assumed to be independent and gene-gene interactions are not taken into account. In biological responses, gene-gene interactions will take place because these responses often result from coregulation of genes [11, 12]. Consequently, by testing each gene independently, weak to small genetic effects

that only in interaction make an important distinction between different study groups will not be detected by using a univariate test.

Gene selection

For functional interpretation the total ranked gene set can be used, but this will include noise, and selection of the most important genes is needed. The difficulty in gene selection is how to define the threshold. The threshold for selecting the differentially expressed genes influences the functional interpretation. Selection of genes is to some extent subjective, because there are no clear thresholds for existing methods. For the t-test, the threshold choice is flexible and the significance level is chosen by the researcher [13, 14]. However, a threshold should preferably be defined in an objective way. Procedures can be applied to correct for multiple testing, such as the family-wise error rate (FWER) or the false discovery rate (FDR) [15, 16]. However, these procedures can be overly stringent, resulting in identification of only the most important changes and possibly discarding other relevant genes [5].

Gene grouping

Each probe on a microarray corresponds to a specific nucleotide sequence, which represents a specific gene. Most genes known to be involved in a functional category are annotated in annotation databases, such as the Gene Ontology (GO) database [17], Kyoto Encyclopedia of Genes and Genomes (KEGG) [18], or Entrez Gene [19]. Whole genome microarrays contain annotated genes as well as non-annotated genes. Although the extent to which spots on whole genome microarrays are annotated has not exactly been established, many known genes are not annotated in functional analysis tools, for example, GO annotated, and are thus lost for biological interpretation when a pathway program uses the GO database as source [19, 20]. However, the non-annotated genes may provide important new targets. Clues on the function of these genes can be obtained by establishing similarities in expression behavior to known genes. Genes with similar gene expression can be identified with self-organizing maps (SOM) and hierarchical clustering [21-23]. SOM has the advantage that it provides an ordering of clusters, whereby each cluster consists of a group of genes with similar gene expression profiles. Grouping based on similarity in expression behavior is also useful for functional interpretation of known genes.

Biological interpretation is the final step in this framework. A useful way to interpret microarray data is pathway analysis. In pathway analysis the effects of treatment on biological processes or coregulated gene sets are studied, rather than effects on individual genes [19, 24]. A commonly used approach is to import a list of genes that meets the threshold criteria into a pathway program, such as the freely available ErmineJ [25], GeneMapp, David/EASE, SAFE [13], or PLAGE [26] or commercially available programs like Metacore [27] or Ingenuity. These programs search through public or private databases to link related genes that are grouped in biological processes.

Recently, new methods have been developed for functional interpretation that circumvent the need to preselect genes [28]. One of these methods is gene set enrichment analysis (GSEA) [29]. This method enables detection of important pathways where all genes in a predefined set (for instance a GO category) change in a coordinated manner [4, 30]. This is highly relevant for studies where subtle, but coordinated changes in expression can be

expected. However, GSEA may have the disadvantage that it is restricted to, and therefore only informs about, functionally annotated genes. Thus not all information that is available in the dataset is used. Nevertheless, the application of GSEA has shown that small effects can be captured when coordinate gene expression changes are taken into account [29].

In this study we describe a framework for functional interpretation of microarray-based expression studies using two real gene expression datasets. For gene ranking and selection, we have examined the usefulness of random forests (RF) [31]. RF is one of the statistical methods that have been developed to select genes from large datasets containing many variables in small sample sizes. RF and other supervised methods like support vector machines (SVM) and discriminant analysis (DA) have mainly been used to select genes that provide the best classification performance for diagnostic purposes (see, e.g. [32, 33]). In microarray studies, RF was shown to outperform other classification methods, especially when the number of classes is moderate [34, 35]. RF could also be a suitable tool to rank and select a larger subset of genes for further interpretation, because it has many advantages [34]. One major advantage of RF is that it provides an importance measure for each gene, which can be used to rank the genes. Furthermore, the advantage of this importance measure is that it takes gene-gene interactions in the ranking of genes into account. In this way, RF is able to capture not only the main effects in a dataset but also the variables with weak to small genetic effects that mainly contribute by interactions with other genes. Interaction between genes increases the importance of the individual interacting genes, making them more likely to be given high importance relative to other genes. Genes with a higher importance index (I_m) are more associated with differences resulting from the treatment. A simulation study has been performed, showing that RF outperforms a univariate method [36]. This study showed that the more interactions are present, the better RF performs compared with a univariate method. Because RF takes gene-gene interactions into account in the ranking of genes, this method was applied within this framework as a tool to rank genes at the first step. However, RF does not provide a threshold to define which genes should be selected for further interpretation. Therefore, after applying RF to rank genes by their I_m, we examined an approach to define a threshold for the genes ranked by RF to select biologically important genes in an objective way. After selection, genes were clustered by SOM, which clusters genes with similar gene expression in ordered profile groups. The advantage of combining results obtained with SOM and information obtained at previous steps is that insight can be obtained as to whether genes within the same profile contribute by their main effect and/or whether interaction effects are present and whether profiles containing relevant biological information are obtained. Finally, for each gene expression dataset, the selected genes obtained by RF were incorporated in pathway programs (Metacore and ErmineJ) and compared with the results obtained with GSEA. Together this provides a stepwise framework focusing on the different considerations in the identification of informative genes and the interpretation of their biological relevance.

Methods

Datasets

To illustrate and examine the framework considerations, we used two whole genome gene expression datasets obtained from the same dietary study. The animal welfare committee of Wageningen University approved the experimental protocol. In this study, two groups of Wistar rats were fed different diets for 2 weeks. One group of rats received a control diet (n=12) and the other an experimental diet (n=12). The experimental diet was identical to the control diet but additionally contained fructo-oligosaccharides. Detailed analysis of the effects of the diet is the subject of another paper. The two datasets were obtained from two different tissues, colon and cecum. RNA from colon mucosa and cecum mucosa was isolated, reverse transcribed into cDNA, labeled, and individually hybridized to Agilent-Whole Rat Genome Microarrays (G4131A). Labeling was performed by incorporating Cy5 for individual samples and Cy3 for pooled RNA. Hybridization and washing were carried out according to Agilent protocols. A total of 24 arrays for colon were analyzed; one array did not pass the quality controls based on MA-plot and signal intensity distribution [9, 37]. Therefore, the colon dataset contained 23 arrays in total. The cecum dataset contained 22 arrays in total, since two cecum RNA samples were excluded because of poor quality of RNA. We preprocessed the microarray datasets as described previously [38]. Only genes with an average signal 1.5 times above the background were taken into account for further data analysis, equal to 28,180 genes for colon and 21,049 genes for cecum. Gene expression values were log-transformed before statistical analyses were performed. The data have been deposited in NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession numbers GSE5943 and GSE8587.

Statistical analyses

t-test. t-tests to obtain t-statistics and corresponding p-values for the differences in mean gene expression between the two treatment groups were performed with the program GeneMaths XT (Applied Math, Sint-Martens-Latem, Belgium). Within the same program FDR analyses according to the Benjamini and Hochberg procedure [15] were performed.

Random forests. In RF a group of tree-based models (the forest) can be used to rank genes with an important contribution to the treatment variable. Each tree starts with the total dataset, which is recursively split into smaller and more homogeneous groups to fit models for predicting the different treatment groups from the selected genes. Within the forest, different trees are obtained by bootstrap sampling and random subset selection. In more detail, each tree is constructed from a bootstrap sample of the total dataset. A bootstrap sample is obtained by sampling observations (e.g., rats) from the original dataset with replacement. The bootstrap sample contains the same number of observations as the original dataset, but some observations are sampled more than once, while others are left out. The sampled observations are used to construct the tree, whereas a class prediction is obtained for each left-out observation, based on its gene expression values. A prediction for the forest is obtained by aggregating the predictions over all trees for which the observation was left out. The prediction error of the forest is then the proportion of misclassified

samples, indicating the performance of the forest to correctly predict the class labels of the different observations. For each split in a tree, the gene that gives the best split is not selected from the total set of genes but from a random subset of genes. The number of randomly selected genes that is used to be searched through for the best split is referred to as m_{try}. RF performance is usually not sensitive to this parameter, and it is suggested to use $\sqrt{total\ nr\ of\ genes}$ as a default value for m_{try} [31, 39]. Comparing the default value and values lower and higher than the default for both colon and cecum, we obtained similar prediction errors for different m_{try} values (data not shown). Therefore, default values for m_{try} ($\sqrt{total\ nr\ of\ genes}$) were chosen for both colon (167 genes) and cecum (145 genes) to perform the RF analyses.

More important genes will discriminate better between the treatment groups and will therefore be present in most of the trees and more often selected at a split close to the total sample. On the other hand, less important genes will be less present in the different trees and selected at splits farther from the total sample. Importance of genes is defined by a measure referred to as the importance index, I_m. For each gene, this I_m is obtained by comparing the predictive performance of the forest for all genes with the predictive performance of the forest in which the values of the gene are randomly permuted in the trees for the left-out observations. Larger differences in the predictive performance give a larger I_m, indicating more important genes. By permuting the values for one gene, not only is the effect of this gene taken into account, but also all possible interactions of this gene with other genes. Interactions between genes increase the I_m for each of the genes that are part of the interaction. In this way, RF takes interactions between genes into account. Several measures of importance are available [39, 40]. To perform the RF analyses we used the scaled mean decrease in classification accuracy. Genes are ranked according to their importance. To obtain stable estimates of the I_m, large numbers of trees in the forest are needed [36, 39]. Also, to capture as many important interactions as possible, huge numbers of trees are required. RF does not overfit; therefore we performed the analyses with a large number of trees (40,000). We used all genes in the dataset in the analysis, and I_m was used as measure to rank the genes.

To obtain a threshold for selection of genes for subsequent interpretation, the permutation test [41, 42] was applied. We used 100 permutation datasets, in which the group labels are randomly permuted. For each permutation dataset, RF analysis was performed with the same parameter settings as for the observed dataset. Next, for each permutation dataset I_m values for the genes were obtained and genes were ranked. The distribution of the I_m values derived from the permutation datasets indicates how the I_m values of the genes behave in the absence of a true association with the treatment. To define the threshold for selecting genes, two approaches were taken. The first approach was to determine the value of I_m where the I_m of the observed dataset was equal to, or lower than, the I_m for at least 1 of the 100 permutation datasets. This corresponds to a significance level of $p<0.01$. The second approach to define the threshold, which is explained and illustrated at the GeneSrF website [43], was to determine the number of genes with I_m larger than the mean value of I_m for the first ranked gene obtained from the 100 permutation datasets. However, this second approach yielded only a small number of genes, 11 for colon and 20 for cecum, with highly stringent p-values of 7×10^{-7} for colon and 9×10^{-6} for cecum. Since this limited number of

genes does not provide sufficient information for pathway analysis, we only used the first approach.

To examine whether RF provides reproducible results over different analyses, we performed several analyses (runs), each time using the same parameter settings but a different seed value. The seed value controls the random number generator, and different seed values generate different forests. The results can be repeated if the same seed value is used. We examined the reproducibility of RF by comparing the I_m of the genes for different runs. Each run can return slightly different results because in RF each tree is constructed on a bootstrap sample of the observations (rats), and at each split of the tree the best discriminating gene is selected from a random subset of genes (m_{try}).

The permutation test that was used to determine the threshold of the I_m was also used to obtain the significance of the prediction error of the RF model. For each permuted dataset, a prediction error was obtained by RF. The proportion of permutation datasets with a prediction error equal to or lower than the prediction error of the RF model of the observed dataset provided the significance of the model.

Software for RF is freely available, including R-packages [39, 44-46] and the original Fortran code [40]. For analyses with RF we have applied the R-package randomForest to obtain the I_m for the different genes.

Gene grouping: SOM

For the gene sets selected with the obtained RF threshold (935 genes in colon, 165 genes in cecum), SOM analyses were performed, in which genes with similar expression are grouped into gene expression profiles. We chose the number of profiles based on the number of genes per profile we expected to be biologically related, and it was therefore set at a mean of 10 genes per profile. This corresponds to 90 SOM profiles for colon and 16 for cecum. To distinguish between genes that mainly contribute by their interaction effect or their main effect, genes selected by RF were compared with the same number of genes ranked by t-test. We explored whether profiles consisting of genes only selected by RF were present, which indicate profiles consisting of gene-gene interaction effects.

To perform SOM analysis, both commercial (e.g., GeneMaths XT) and free open-source (e.g., Orange machine learning software [47] at http://www.ailab.si/orange) are available. In this study we used GeneMaths XT (Applied Math) software packages to obtain the SOM profiles.

Biological interpretation: pathway analysis

For the genes selected by RF, we performed pathway analyses for biological interpretation. The pathway results obtained for genes selected by RF were compared with pathway results obtained for the same number of genes selected by t-test, to ensure comparability. For pathway analysis we used the freely available software ErmineJ [25] and the commercial program Metacore [27]. ErmineJ is a web-based application for identification of GO processes on input gene sets. Metacore is a package of GeneGo (St. Joseph, MI).

In ErmineJ we used overrepresentation analysis (ORA); in Metacore GO processes were used for pathway analysis. For both ErmineJ ORA analysis and Metacore GO processes, gene sets existing of 5–250 genes were tested. In both analyses, gene lists selected by RF or

t-test were classified into GO processes. These processes were ranked according to their p-value, which represents the probability that a particular process is selected by chance. Each pathway program uses different statistical tests to calculate these probabilities; this issue is beyond the scope of this paper and is discussed by others [19, 48, 49]. For both programs we selected pathways with two selection criteria: 1) the pathways should have a $p<0.001$, and 2) the pathways should include at least three selected genes.

We also analyzed which biological pathways were enriched with GSEA [29]. In GSEA, enrichment of genes in a gene set is based on the ranking of the genes within the whole dataset [28]. We included functional c2 gene sets originated from KEGG, GenMapp, and BioCarta with 5–500 genes with FDR q-value<0.25 and ranked on normalized enrichment score (NES) and nominal p-value.

Results

Whole genome arrays are not fully examined in pathway analysis programs

Whole genome microarray analysis combined with pathway analysis is an attractive approach to identification of the effects of an intervention, but the analysis is limited to those genes that are annotated in the database used by the program. To assess completeness of annotation of whole genome arrays in pathway programs, we examined first the extent to which genes were incorporated in the analysis in three different pathway programs, Metacore (GeneGo), ErmineJ, and GSEA. This was performed for the two most widely used array platforms, Agilent and Affymetrix, and for three different species, human, mouse, and rat. Only 23–48% of the probes on whole genome microarrays are translated to functional categories by these programs (table 1). ErmineJ was not included because it does not provide the number of incorporated genes. Annotation in this program is based on the specific GO term(s) linked to the gene, which for the Agilent 44K rat array applies to 7,437 genes (18%). Altogether, analysis only based on functional annotation and co-occurrence in gene sets leaves out at least half of the microarray data, and thereby potential new targets.

Table 1: Percentage of probes from whole genome microarrays identified by the pathway programs Metacore and GSEA.

Number of probes imported		Number of probes linked to program database			
		Metacore*		GSEA†	
		Number	Percentage	Number	Percentage
Agilent					
Human	41675	12976	31	17517	42
Mouse	41534	13714	33	19589	47
Rat	41372	9489	23	14631	35
Affymetrix					
Human	54675	22792	42	20606	38
Mouse	45102	18105	40	21891	48
Rat	44761	12259	39	13342	43

* Spots linked to a GO-term

† Spots linked to a gene symbol

Information content of gene expression datasets

In both gene expression datasets (p=28,180 for colon, p=21,049 for cecum) the extent of differential gene expression induced by the dietary treatment was small: in colon, 179 genes were differentially expressed with a change of >1.5-fold, while in cecum the number of differentially expressed genes was 164. Based on fold change the two datasets are similar in number of expressed genes and magnitude of differential expression (fold change). However, the datasets differed in the significance of expression, with the colon dataset containing substantially more significantly differentially expressed genes (table 2). With a t-test threshold of $p<0.001$, 803 genes were differentially expressed in the colon dataset, while 123 genes were differentially expressed in cecum. Application of FDR using a threshold of $q<0.01$ resulted in selection of 231 genes in colon and 19 genes in cecum. RF models were found to be significant in both colon ($p<0.02$) and cecum ($p<0.01$), indicating that gene expression differences were present.

Table 2: Characteristics of the colon and cecum dataset.

Dataset	Total number of genes in dataset	Fold change >1.5 fold*	t-test p<0.001	FDR q<0.01
Colon	28180	164	803	231
Cecum	21049	179	123	19

*Fold change experimental diet/control diet.

Gene ranking: taking gene-gene interactions into account

Genes were ranked according to their I_m obtained by RF. To obtain insight into the ranking of genes by RF, we compared the results from RF with the ranking of genes by the commonly used t-test. For the genes present in the dataset the absolute values for the t-statistics are plotted against the I_m of RF (see figure 1). In both datasets, I_m obtained from RF does show a similar trend with t-statistics. Both RF and t-test rank genes in common (figure 1, Box A), indicating strong gene effects related to the treatment. Genes ranked high by RF, compared with t-test (figure 1, Box B), are indicative of weak gene effects that are likely to be related to the treatment in interaction with other genes.

Gene selection: defining an objective threshold

We aimed to define an objective threshold for I_m by using a permutation approach (see Methods). This permutation test provides an indication of where noise starts to interfere with real gene effects. For both colon and cecum the highest-ranked genes from the observed dataset had I_m values higher than the ranked I_m values obtained from the permuted datasets (see figure 2). To define the threshold, we determined the I_m value where genes in the observed dataset have equal or higher I_m values relative to the genes in the permuted datasets. The point at which the I_m values of the observed dataset equaled that of at least 1 of the 100 permutated datasets was chosen as threshold, which is equal to a significance level for the I_m of $p<0.01$.

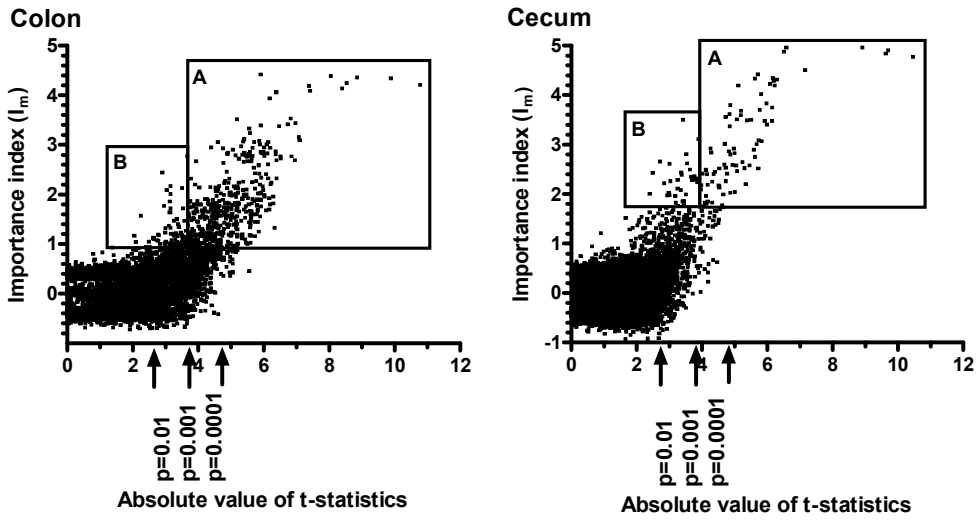


Figure 1: Plot of absolute value of t-statistics against I_m for colon (left) and cecum (right) dataset. Box A: Genes highly ranked by both RF and t-test. Box B: Genes highly ranked exclusively by RF.

We performed 15 runs (each with a different seed value) resulting in very similar thresholds (results not shown). For colon a mean threshold of I_m=0.906 and for cecum a mean threshold of I_m=1.753 were obtained. For each run, the genes with I_m values above the threshold were determined. Genes with higher I_m values were always selected over the different runs. However, genes with ranking close to the threshold (lower I_m values) were not selected over all runs; thus the selection of these genes varied between different runs. We chose to include all genes that were selected in at least one run and not only the overlapping genes, because the number of genes that were additionally selected over increasing numbers of runs decreased rapidly (table 3; figures 3A and B, for colon and cecum, respectively).

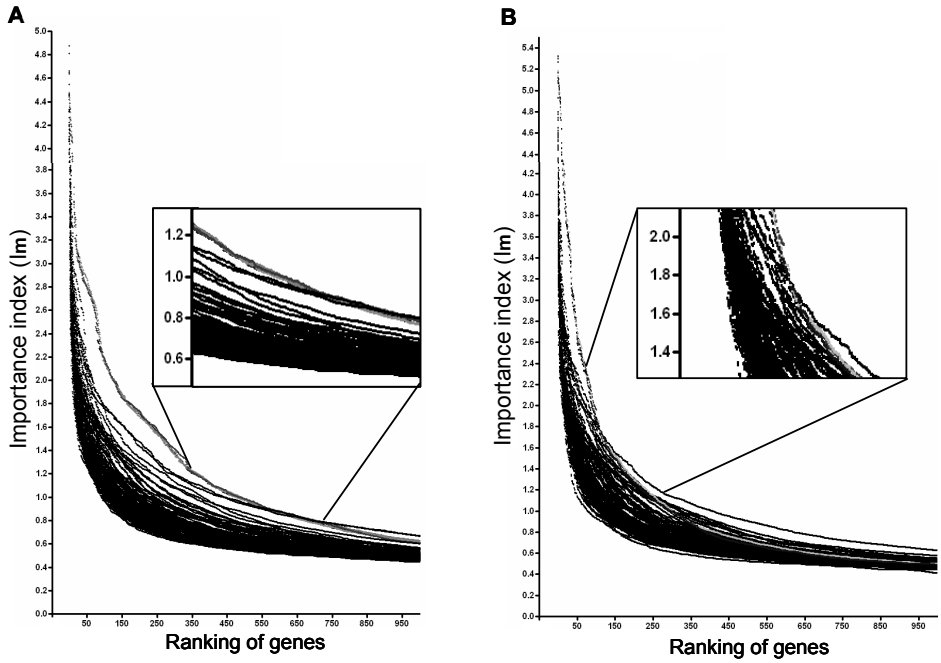


Figure 2A and 2B: Genes, of 100 random sets (black lines) and real sets with different seed values (colored lines), ranked by the I_m values. For colon (A) and cecum (B) datasets. **For full color figure, see page 165.**

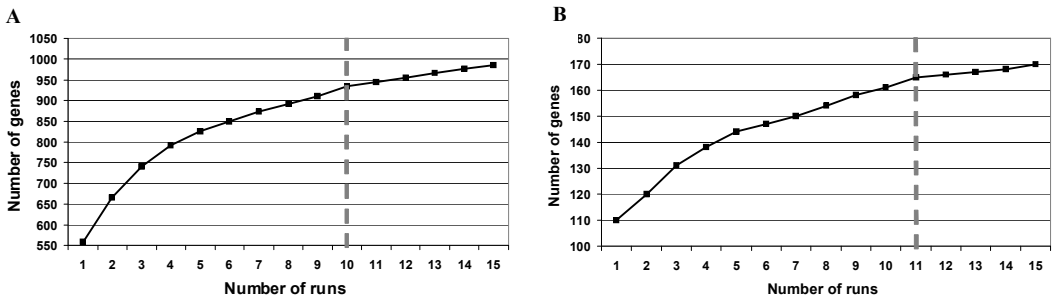


Figure 3A and 3B: Genes selected by RF thresholds Im> 0.906 for colon (A) and Im>1.753 for cecum (B). The total number of selected genes is plotted against the number of runs.

This likely indicates that additionally selected genes are truly affected by the treatment and not randomly selected noise. After 10 runs for colon and 11 runs for cecum, the proportion of genes additionally selected became and remained <2%. Therefore, more runs were not performed. Combining the results of different runs resulted in the selection of 935 genes above the threshold for colon and 165 genes above the threshold for cecum. These genes were selected as the set of genes being related to the treatment.

Table 3: Selection of genes by RF threshold.

A: Colon

Run	Number of genes*	Total number of genes†	Genes added	
			Number	Percentage
1	558	558	-	-
2	552	665	107	19.2
3	558	740	75	11.3
4	558	791	51	6.9
5	557	825	34	4.3
6	562	849	24	2.9
7	542	873	24	2.8
8	557	891	18	2.1
9	564	911	20	2.2
10	549	**935**	24	2.6
11	560	945	10	1.1
12	554	955	10	1.1
13	540	966	11	1.2
14	573	977	11	1.1
15	547	985	8	0.8

* Number of genes selected with threshold I_m>0.906.

† The number of genes selected after each additional run.

B: Cecum

Run	Number of genes*	Total number of genes†	Genes added	
			Number	Percentage
1	110	110	-	-
2	109	120	10	9.1
3	118	131	11	9.2
4	112	138	7	5.3
5	111	144	6	4.3
6	108	147	3	2.1
7	112	150	3	2.0
8	108	154	4	2.7
9	112	158	4	2.6
10	115	161	3	1.9
11	111	**165**	4	2.5
12	114	166	1	0.6
13	108	167	1	0.6
14	115	168	1	0.6
15	108	170	2	1.2

* Number of genes selected with threshold I_m>1.753.

† The number of genes selected after each additional run.

Comparison of gene selection by RF, t-test, and fold change

Genes selected based on the RF threshold (935 genes in colon and 165 genes in cecum) were compared with an equal number of genes selected by t-test. For t-test this resulted in inclusion of genes with $p<0.0014$ ($q<0.04$) for colon and $p<0.0018$ ($q<0.23$) for cecum. In colon 679 genes (72.6%) and in cecum 112 genes (67.9%) overlapped between RF and t-test. As shown in the volcano plots (figure 4), gene sets selected by RF include the most significant genes based on t-test, as was also seen in figure 1. Furthermore, the volcano plots show that RF and t-test also differ in selection of genes. Several genes with high fold change, which would not have been selected based on t-test alone, are also selected by RF.

For both datasets, the set of selected genes by RF were used for subsequent gene grouping and biological interpretation.

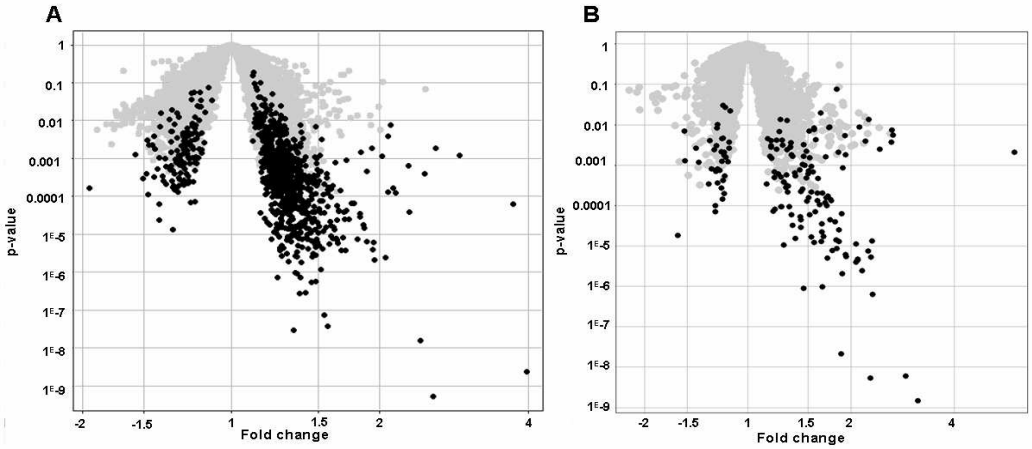


Figure 4: Volcano plots for colon (A) and cecum (B). Fold change is plotted against p-value. All genes are shown, genes selected by RF are shown in black (935 for colon, 165 for cecum).

Gene grouping: obtaining gene expression profiles by SOM

For grouping of the genes selected by RF, we applied SOM analysis to find groups of highly correlated genes. While SOM is mostly used to identify patterns in time or as a result of multiple treatments [22], it will also identify patterns of coordinate changes over a number of animals. In figure 5, A and B, the groups of genes with similar expression are shown for colon and cecum, respectively. For both colon and cecum, profiles are present that consist mainly of genes that are selected exclusively by RF (light gray in figure 5). SOM analyses for genes selected by the t-test did not result in profiles consisting of genes exclusively selected by t-test (data not shown). Apparently, RF selects genes with main effects similarly to the t-test, but additionally selects genes (not selected by t-test) that can be grouped in profiles, which are likely to be related to the treatment by gene-gene interaction effects.

We examined whether the genes exclusively selected by RF and highly enriched within one profile shared similar biological functions. Therefore we selected profiles consisting of mainly RF-selected genes. For colon two profiles and for cecum one profile was selected (figure 5, white boxes).

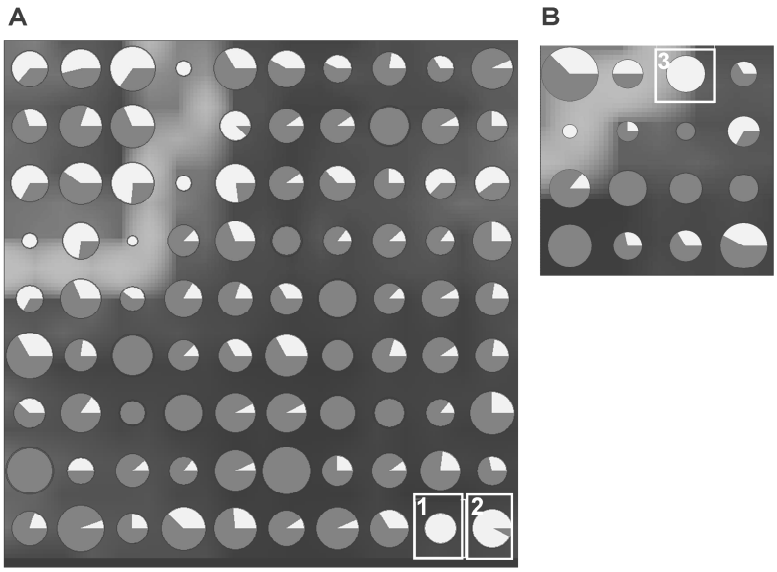


Figure 5: SOM profiles for colon (A) and cecum (B). The total number of SOM-profiles was arbitrarily set to 90 for colon and 16 for cecum, corresponding to an average of approximately 10 genes per profile. The size of the circles corresponds to the number of genes included in the group (range of genes per profile: colon 1-19, cecum 2-27). Within each profile, genes that overlap between RF and t-test are shown in dark gray, and genes exclusively selected by RF are shown in light gray. Genes in profiles 1, 2 and 3 were analyzed in more detail.

The first colon profile (profile 1) consisted of nine genes, four genes with unknown function expressed sequence tags (ESTs) and five genes that were annotated but not classified to a known GO process. After literature and database search these five genes could not be linked to a single biological process (table 4). The second colon profile (profile 2) consisted of 13 genes, of which 12 were only selected by RF. Five genes were annotated in a GO process (bold gene names in table 4), of which four are part of the same GO process: cellular component organization and biogenesis. The remaining eight genes consisted of two ESTs and six genes that are presently poorly understood, because further database and literature mining did not reveal a relation to a known biological process. One of these six (palladin) was recognized to play a role in maintaining normal actin cytoskeleton architecture [50], indicating a possible role in the same biological process as the four annotated genes within this SOM profile.

The cecum profile consisting of exclusively RF-selected genes (profile 3) consisted of 13 genes, comprising 10 unique genes. Three of the 10 genes were annotated by GO, of which 2 are part of the GO process immune response. Further database and literature mining revealed that six of the seven other genes had a function related to immune response (table 4). This confirms the notion that genes with a similar expression profile selected from a microarray dataset exclusively by RF may be enriched in the same biological process. It further indicates that this is a strategy to hunt for biological function of genes and to reveal new biological processes related to treatment.

Table 4: Genes mainly selected exclusively by RF, grouped in SOM profiles (white boxes figure 5).
A: Colon SOM profile number 1.

Gene name*	Sequence ID	Gene symbol†	FC‡	p-value	q-value
3222401M22Rik protein (LOC363231)	XM_343571	-	1.39	0.005	0.08
2410024A21Rik protein (LOC314415)	XM_234506	-	1.21	0.014	0.13
Rattus norvegicus cDNA clone UI-R-A1-dv-f-02-0-UI 5'	BF558849	-	1.20	0.007	0.10
Uronyl-2-sulfotransferase	XM_341728	*Ust*	1.12	0.074	0.31
Ring finger protein 10	XM_213797	*Rnf10*	1.21	0.026	0.18
Midnolin	TC480469	*Midn*	1.31	0.002	0.05
Mitsugumin 29	XM_342316	*Mg29*	1.16	0.012	0.13
Carbonic anhydrase I (Carbonate dehydratase I)	XM_226922	*Ca1*	1.28	0.007	0.10
Polyglutamine-containing protein	BF546374	-	1.22	0.001	0.04

* None of these genes were annotated by GO.
† Genes without official gene symbol are indicated with -.
‡ Fold change experimental diet/control diet.

B: Colon SOM profile number 2.

Gene name*	Sequence ID	Gene symbol†	FC‡	p-value	q-value
Hypothetical protein FLJ32871	XM_219819	-	1.26	0.025	0.18
GCD14/PCMT domain containing protein	NM_001007706	-	1.23	0.015	0.14
Telomeric repeat binding factor 2	XM_341683	*Terf2*	1.26	0.002	0.05
Probable nocturnin protein	XM_344988	-	1.11	0.184	0.48
cGMP-dependent protein kinase 1, beta isozyme	XM_219807	*Prkg1*	1.21	0.010	0.11
High mobility group nucleosomal binding domain 1	BI303604	*Hmgn1*	1.11	0.156	0.44
Cyclin-dependent kinase 5	NM_080885	*Cdk5*	1.18	0.021	0.16
Beta-sarcoglycan	XM_223355	*Sgcb*	1.11	0.096	0.35
Phosphodiesterase isoform	AF053097	*Pde*	1.18	0.005	0.08

Table 4B (continued)

Gene name*	Sequence ID	Gene symbol†	FC‡	p-value	q-value
Palladin, cytoskeletal associated protein	XM_214338	*Palld*	1.20	0.010	0.11
splA/ryanodine receptor domain and SOCS box containing 3	XM_220230	*Spsb3*	1.20	0.003	0.07
Kinesin family member 5B	XM_341538	*Kif5b*	1.24	0.001	0.03
Acyl Transferase	XM_235527	*Mct*	1.13	0.010	0.11

* Genes annotated by GO are presented in bold.
† Genes without official gene symbol are indicated with -.
‡ Fold change experimental diet/control diet.

C: Cecum SOM profile number 3.

Gene name*	Sequence ID	Gene symbol†	FC‡	p-value	q-value
Anti-NGF30 antibody light-chain , variable and constant regions	U39609	-	2.65	0.005	0.31
Ig germline kappa-chain gene C-region	M12981	*Igkc*	2.63	0.007	0.34
Immunoglobulin joining chain	XM_341195	*Igj*	2.11	0.009	0.35
Immunoglobulin rearranged κ-chain mRNA variable (V) region	CO562777	*Igkv*	1.92	0.005	0.31
Anti-acetylcholine receptor antibody gene, κ-chain, VJC region	L22655	-	2.25	0.013	0.39
Ig germline kappa light chain joining (J) segments	J00746	*Igkjca*	1.72	0.009	0.35
Periostin, osteoblast specific factor	XM_342245	*Postn*	1.93	0.002	0.23
Immunoglobulin kappa light chain variable region	AF217591	*Igkv*	1.73	0.009	0.35
Chemokine (C-X-C motif) ligand 12	NM_022177	*Cxcl12*	1.63	0.020	0.46
Ig active kappa-chain mRNA VJ-region from immunocytoma IR162	M15402	*Igkac*	2.20	0.004	0.28
		Igkac	1.56	0.008	0.35
		Igkac	1.56	0.008	0.34
		Igkac	1.58	0.004	0.29

* Genes annotated by GO are presented in bold.
† Genes without official gene symbol are indicated with -.
‡ Fold change experimental diet/control diet.

Biological interpretation: pathway analysis to obtain biological processes

To examine whether pathway programs are able to identify differences between RF-selected genes and t-test-selected genes, we applied pathway analysis for the set of genes selected by RF and compared this with the same number of genes selected by t-test (935 genes for the colon dataset and 165 for the cecum dataset). To ensure that we covered different pathway analysis methods, we used two pathway programs, Metacore and ErmineJ. For both colon (table 5) and cecum (table 6) the comparison between RF- and t-test-based selection showed highly comparable results per pathway program. However, the ranking of processes was somewhat different, and each selection method (RF or t-test) identified some unique processes.

GSEA does not require preselection of genes, although information may be lost because of incomplete annotation. GSEA is especially suited to identifying processes based on interaction. To see whether similar or complementary information is obtained, we analyzed the complete colon and cecum datasets with GSEA. We focused on pathway-related GSEA gene sets, obtained from GO, GenMapp, and Biocarta, to allow for comparison. Only a few gene sets were found to be significantly enriched (FDR<0.25 according to GSEA): 12 in colon and 6 in cecum. The small number of processes identified by GSEA analysis suggests that information is lost. The program does give some overlapping pathways in colon, but in cecum other processes are selected. In both cases no overlap with processes only selected with RF was found.

Discussion

We described a framework for physiological interpretation of gene expression data. This framework (see Box 1) consists of the following steps: genes are first ranked, the relevant genes are selected, and the selected genes are grouped according to their expression profile and then biologically interpreted. The considerations underlying the different steps are illustrated with two real gene expression datasets. We show several features of RF that should be part of any data analysis framework. These are 1) all genes in the dataset are included in the analysis, 2) interaction between genes is taken into account, and 3) a well-defined gene set can be selected by using an objective threshold.

For human, mouse, and rat whole genome arrays, the number of annotated genes is less than half of the genes present on the array. Consequently, analysis only based on functional annotation and co-occurrence in gene sets filters out half of the information present in the microarray dataset. Well-studied biological processes are better represented in pathway databases [19]. Therefore, conclusions obtained from data analysis based only on pathway programs are biased toward the well-annotated biological processes. By including all genes from a whole genome dataset, it is possible to find genes or processes less defined in databases but could be attractive new targets for drug development or nutritional intervention. For both colon and cecum, genes exclusively selected within one SOM profile belonged to the same biological process: cellular component organization and biogenesis (colon) and immune response (cecum), respectively. Because only a few genes within these profiles were GO annotated, these processes were not selected by the different pathway programs. By literature and database search we could clearly identify these genes as part of this process.

Table 5: Biological processes in the colon dataset selected by Metacore, ErmineJ and GSEA.

A: Metacore and GSEA

t-test*	p-value	RF*	p-value	GSEA genesets†	NES	p-value
mitochondrial electron transport, NADH to ubiquinone	3E-10	**mitochondrial electron transport, NADH to ubiquinone**	6E-08	***Mitochondria***	2.09	0.00
oxidative phosphorylation	4E-09	**protein targeting to mitochondrion**	1E-07	***Electron transport chain***	1.97	0.00
organelle ATP synthesis coupled electron transport	5E-09	**mitochondrial transport**	2E-07	***Oxidative phosphorylation***	1.96	0.00
ATP synthesis coupled electron transport	5E-09	**oxidative phosphorylation**	3E-07	*Propanoate metabolism*	1.92	0.00
protein targeting to mitochondrion	1E-08	**electron transport**	3E-07	*Proteasome degradation*	1.86	0.01
mitochondrial transport	2E-08	**organelle ATP synthesis coupled electron transport**	5E-07	*Proteasome*	1.83	0.01
electron transport	2E-06	**ATP synthesis coupled electron transport**	5E-07	*Free Radical Induced Apoptosis*	1.80	0.01
regulation of carbohydrate metabolic process	2E-05	**regulation of carbohydrate metabolic process**	2E-05	*Butanoate metabolism*	1.79	0.00
muscle filament sliding	1E-04	**coenzyme metabolic process**	2E-05	***Tricarboxilic acid cycle***	1.78	0.00
coenzyme metabolic process	1E-04	**energy derivation by oxidation of organic compounds**	1E-04	*Programmed cell death*	1.77	0.00
regulation of insulin secretion	2E-04	**regulation of insulin secretion**	1E-04	*Valine leucine and isoleucine degradation*	1.75	0.02
main pathways of carbohydrate metabolic process	2E-04	cofactor metabolic process	2E-04			
biopolymer catabolic process	2E-04	response to inorganic substance	2E-04			
response to copper ion	2E-04	**response to copper ion**	2E-04			

Table 5A (continued)

t-test*	p-value	RF*	p-value	GSEA genesets†	NES	p-value
nucleosome assembly	3E-04	**nucleosome assembly**	3E-04			
feeding behavior	4E-04	**insulin secretion**	3E-04			
insulin secretion	4E-04	regulation of secretion	4E-04			
chromatin assembly	4E-04	aerobic respiration	6E-04			
energy derivation by oxidation of organic compounds	4E-04	response to metal ion	6E-04			
monocarboxylic acid metabolic process	6E-04	**main pathways of carbohydrate metabolic process**	8E-04			
response to toxin	9E-04	**chromatin assembly or disassembly**	8E-04			
		response to toxin	8E-04			
		positive regulation of pseudopodium formation	8E-04			
		regulation of hormone secretion	9E-04			
		peptide hormone secretion	9E-04			
		peptide secretion	9E-04			
		protein targeting	9E-04			

* Gene subsets of t-test and RF were used as input for Metacore. Overlapping processes between the two genesets (t-test and RF) are presented in bold.

† For GSEA the whole dataset was used, only the genesets compiled from publicly available databases are included.

B: ErmineJ and GSEA

t-test*	p-value	RF*	p-value	GSEA genesets†	NES	p-value
Mitochondrial electron transport, NADH to ubiquinone	5E-18	**mitochondrial electron transport, NADH to ubiquinone**	2E-14	***Mitochondria***	2.09	0.00
ATP synthesis coupled electron transport (sensu Eukaryota)	3E-15	**ATP synthesis coupled electron transport (sensu Eukaryota)**	6E-12	***Electron transport chain***	1.97	0.00
protein biosynthesis	5E-14	**aerobic respiration**	2E-11	***Oxidative phosphorylation***	1.96	0.00
Macromolecule biosynthesis	1E-13	**tricarboxylic acid cycle**	7E-09	*Propanoate metabolism*	1.92	0.00
aerobic respiration	2E-11	**protein targeting to mitochondrion**	1E-08	*Proteasome degradation*	1.86	0.01
Tricarboxylic acid cycle	2E-10	**main pathways of carbohydrate metabolism**	5E-08	*Proteasome*	1.83	0.01
main pathways of carbohydrate metabolism	4E-09	**acetyl-CoA catabolism**	2E-07	*Free Radical Induced Apoptosis*	1.80	0.01
acetyl-CoA catabolism	7E-09	**protein biosynthesis**	4E-07	*Butanoate metabolism*	1.79	0.00
protein targeting to mitochondrion	1E-08	**generation of precursor metabolites and energy**	3E-06	***Tricarboxilic acid cycle***	1.78	0.00
generation of precursor metabolites and energy	1E-07	**proton transport**	6E-06	*Programmed cell death*	1.77	0.00
proton transport	5E-07	**macromolecule biosynthesis**	6E-06	*Valine leucine and isoleucine degradation*	1.75	0.02
pyruvate metabolism	2E-06	**oxidative phosphorylation**	2E-05			
hexose biosynthesis	5E-06	**pyruvate metabolism**	2E-05			
oxidative phosphorylation	2E-05	**DNA fragmentation during apoptosis**	4E-05			
hydrogen transport	2E-05	**ATP biosynthesis**	6E-05			

Table 5B (continued)

t-test*	p-value	RF*	p-value	GSEA genesets†	NES	p-value
Establishment and/or maintenance of chromatin architecture	2E-05	cellular respiration	7E-05			
DNA fragmentation during apoptosis	4E-05	**DNA catabolism**	7E-05			
fatty acid beta-oxidation	5E-05	**disassembly of cell structures during apoptosis**	7E-05			
ATP biosynthesis	6E-05	**ATP synthesis coupled proton transport**	7E-05			
DNA catabolism	7E-05	**coenzyme metabolism**	8E-05			
Disassembly of cell structures during apoptosis	7E-05	**hexose biosynthesis**	9E-05			
ATP synthesis coupled proton transport	7E-05	**protein secretion**	9E-05			
protein secretion	9E-05	tricarboxylic acid cycle intermediate metabolism	1E-04			
cellular metabolism	1E-04	**apoptotic nuclear changes**	1E-04			
apoptotic nuclear changes	1E-04	**sensory perception of sound**	2E-04			
DNA packaging	1E-04	**hydrogen transport**	2E-04			
sensory perception of sound	2E-04	**acyl-CoA metabolism**	2E-04			
acyl-CoA metabolism	2E-04	**purine ribonucleoside triphosphate biosynthesis**	2E-04			
purine ribonucleoside triphosphate biosynthesis	2E-04	**electron transport**	3E-04			
electron transport	3E-04	carbohydrate metabolism	9E-04			
protein targeting	3E-04					
fatty acid oxidation	4E-04					
protein amino acid deacetylation	5E-04					
fatty acid metabolism	5E-04					

Table 5B (continued)

t-test*	p-value	RF*	p-value	GSEA genesets†	NES	p-value
Establishment of protein localization	7E-04					
coenzyme metabolism	8E-04					
oxygen and reactive oxygen species metabolism	9E-04					

*Gene subsets of t-test and RF were used as input for ErmineJ. Overlapping processes between the two genesets (t-test and RF) are presented in bold.

† For GSEA the whole dataset was used, only the genesets compiled from publicly available databases are included.

Table 6: Biological processes in the cecum dataset selected by Metacore, ErmineJ and GSEA.
A: Metacore and GSEA

t-test *	p-value	RF*	p-value	GSEA genesets†	NES	p-value
feeding behavior	2E-08	**feeding behavior**	6E-10	*Cell cycle regulator*	1.75	0.01
regulation of insulin secretion	4E-08	leading edge cell differentiation	1E-08	*Cholesteron biosynthesis*	1.67	0.02
regulation of carbohydrate metabolic process	3E-07	**regulation of insulin secretion**	3E-08	*cell proliferation*	1.66	0.00
insulin secretion	9E-07	**eating behavior**	9E-08	*Interleukin 10 pathway*	1.61	0.03
regulation of hormone secretion	3E-06	**regulation of carbohydrate metabolic process**	3E-07	*Caspase cascade*	1.56	0.04
peptide hormone secretion	3E-06	**insulin secretion**	7E-07	*Proliferation*	1.56	0.00
peptide secretion	3E-06	**regulation of hormone secretion**	2E-06			
eating behavior	6E-06	**peptide hormone secretion**	2E-06			
peptide transport	9E-06	**peptide secretion**	2E-06			
regulation of lipid metabolic process	5E-05	**peptide transport**	7E-06			
hormone secretion	5E-05	epithelial cell differentiation	2E-05			
regulation of angiogenesis	2E-04	**regulation of secretion**	3E-05			
regulation of secretion	4E-04	**hormone secretion**	4E-05			
generation of a signal involved in cell-cell signaling	5E-04	**regulation of lipid metabolic process**	4E-05			
monocarboxylic acid transport	5E-04	morphogenesis of an epithelium	1E-04			
		cellular defense response	3E-04			
		generation of a signal involved in cell-cell signaling	4E-04			

*Gene subsets of t-test and RF were used as input for Metacore. Overlapping processes between the two genesets (t-test and RF) are presented in bold.
† For GSEA the whole dataset was used, only the genesets compiled from publicly available databases are included.

B: Ermine J and GSEA

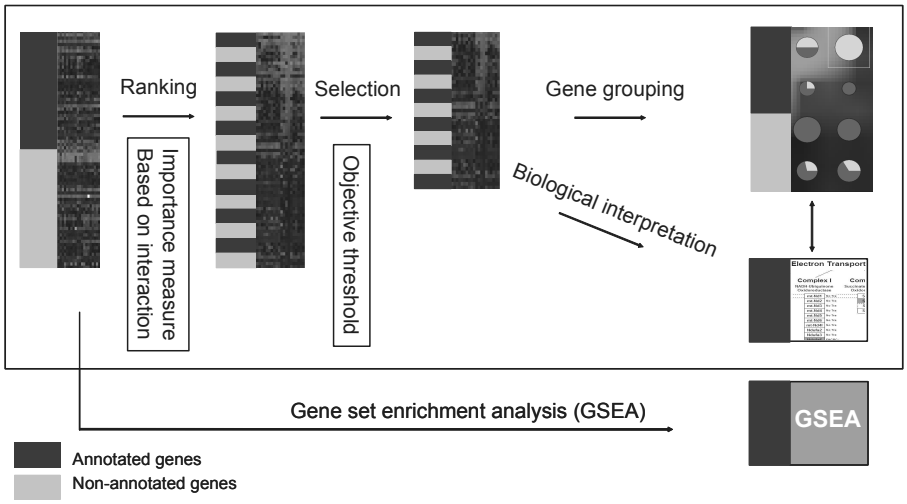
t-test*	p-value	RF*	p-value	GSEA genesets†	NES	p-value
digestion	2E-06	**cellular defense response**	1E-06	*Cell cycle regulator*	1.75	0.01
regulation of angiogenesis	6E-06	epithelial cell differentiation	1E-05	*Cholesteron biosynthesis*	1.67	0.02
cellular defense response	4E-05	**oxygen and reactive oxygen species metabolism**	3E-05	*cell proliferation*	1.66	0.00
muscle development	6E-05	neuron migration	3E-05	*Interleukin 10 pathway*	1.61	0.03
wound healing	3E-04	T cell activation	4E-05	*Caspase cascade*	1.56	0.04
actin cytoskeleton organization and biogenesis	4E-04	response to wounding	5E-05	*Proliferation*	1.56	0.00
oxygen and reactive oxygen species metabolism	5E-04	**digestion**	5E-05			
tissue development	6E-04	defense response	1E-04			
regulation of cell differentiation	9E-04	cell migration	2E-04			
		wound healing	3E-04			
		cell motility	7E-04			
		regulation of cell differentiation	8E-04			
		neurogenesis	9E-04			

*Gene subsets of t-test and RF were used as input for ErmineJ. Overlapping processes between the two genesets (t-test and RF) are presented in bold.

† For GSEA the whole dataset was used, only the genesets compiled from publicly available databases are included.

A major strength of whole genome microarray studies is that the expression levels of all genes are displayed, allowing for identification of gene-gene interactions. RF was chosen to rank genes because its measure of importance takes possible interactions between genes into account. Compared with the results obtained by t-test, RF selected genes with main effects but additionally was able to capture weak effects. In studies with small gene expression changes that are not significant independently but occurring in one group may be of large relevance, this is an advantage. For example, it enables identification of possible side effects in drug studies or expected subtle differences in nutritional studies. In our study, application of RF in combination with SOM indeed showed enriched profiles containing mainly genes selected exclusively by RF and not by t-test. Genes within these profiles are therefore contributing by gene-gene interactions.

Box 1: A framework for identification of physiological responses in microarray based gene expression studies. The framework is composed of the following steps: Gene ranking, gene selection, gene grouping and biological interpretation. Essential features of the data-analysis framework are that 1) all genes (annotated and non-annotated) in the dataset are included in the analysis, 2) interaction between genes is taken into account and, 3) an objective threshold is used for selection of a well-defined gene set. Random forest has these features. Gene grouping can provide information on new targets and add information above pathway analysis. Despite loss of information due to incomplete annotation of the complete dataset, Gene set enrichment analysis can provide additional information on related genes with small differences.



By applying a permutation test, we defined a threshold for RF to select genes in an objective way. Comparison of different runs showed that the most important genes were consistently selected. However, selection of genes ranked closely above the threshold varied between different runs. We chose to include genes that were additionally selected over different runs in the total selected gene set. By including genes selected additionally by different runs there is a chance that more false positives were included in the selection. If we would have chosen to select the set of genes that overlapped in all runs, we might discard truly relevant genes (false negatives). We reasoned that the increased information available for pathway analysis outweighed the potential disadvantage of including some noise, especially since in dietary studies gene expression changes of interest are usually small. Furthermore, the results show that the number of additionally selected genes decreased rapidly for each additional run. Because there was large overlap, it is less likely that many of the additionally selected genes were noise. Thus, within this framework, RF is a useful tool to select a well-defined set of genes for further interpretation.

SOM was applied to find groups of genes with similar gene expression profiles. Other approaches to find gene groups, such as hierarchical clustering, can be used with the same objective [21]. However, SOM has the advantage, compared with other clustering methods, that it provides an ordering of the profiles. While individual genes may have small gene expression differences, groups of similarly behaving genes can be biologically significant. When SOM analysis is applied to whole genome datasets, unrelated data will also produce clusters, without any physiological relevance [21]. This can be overcome by selecting a subset of genes and examining whether biological valid clusters are obtained. The number of clusters is specified by the user. Specifying larger and smaller numbers of profiles within a certain range does not impact the interpretation of the results, since SOM provides an ordering in the profiles. For both colon and cecum, genes selected by RF and analyzed by SOM provided profiles consisting of genes with similar biological function. In the colon dataset, a SOM profile consisted of genes belonging to the same GO process and genes with poorly identified functions. This could be a starting point to identify possible biological function of the non-identified genes. Using SOM within this framework can provide information on genes with unknown function and help to identify biological processes not captured by pathway analyses. Therefore SOM is a useful tool for identification of biological processes in addition to pathway analysis.

The pathway analysis based on the subset of genes obtained by RF and t-test shows overlap for the selected processes; however, different processes were additionally obtained by RF. Remarkably, GSEA only returned a few gene sets connected to public databases that were significantly enriched in colon or in cecum. The small number of processes identified by GSEA analysis suggests that information is lost. On the other hand, GSEA did provide biological processes not found in the other pathway programs. Although only a few processes were found by GSEA, these are worth exploring because they may consist of related genes with small differences. Thus, in the context of the framework discussed in this paper, GSEA may additionally be applied.

The advantage of this framework is that different methods can be applied at different steps, depending on the aim and preferences of the researcher. For example, other methods that take interactions into account could be used instead of RF. A next step is to extensively compare different methods that take gene-gene interaction into account to select

biologically relevant genes. There are several advantages to the use of RF within this framework to rank and select genes. In a previous simulation study, Lunetta et al. [36] showed that the more interactions that are present in the dataset, the more RF outperforms a univariate test-statistic in prioritizing the important variables. In our study we used two real datasets with subtle gene expression changes and showed that RF in combination with SOM can be used to extract a biologically meaningful group of genes, such as the set of immune response genes in the cecum dataset that would be discarded with univariate tests such as the t-test. As mentioned above, it returns an importance factor for each gene (I_m) in which gene-gene interactions are taken into account. On the basis of this I_m, we showed an approach that can be used to define an objective threshold for selection of genes. Besides two classes, RF can also be applied to multi-class problems. Furthermore, free software is available for RF whereby only a few parameters need to be defined [39]. Also, users can easily obtain a gene list for further interpretation without the need to understand the finer details of the method thoroughly. Therefore, within this framework RF is a suitable and practical tool to rank and select genes. Combined with gene grouping by SOM and pathway programs, this framework is helpful to obtain insight in the biological processes. These physiological effects are the main focus for further confirmatory and mechanistic studies.

In conclusion, in this study we have examined the application of a framework in which all genes in a microarray dataset are analyzed. Within this framework, application of RF has the advantage that it takes gene-gene interactions in the ranking of genes into account. Also, selection of genes by an objective threshold provides a well-defined set of genes for further interpretation. Groups of genes within this set are identified by SOM analysis. In combination with pathway analyses it provides valuable information on biological processes involved in the treatment.

Acknowledgements

The authors thank Dr. E. M. van Schothorst from RIKILT Institute of Food Safety, Dr. P. Wang from Maastricht University, and Dr. D. Molenaar from NIZO Food Research for helpful comments and suggestions. We also thank Dr. T. Travis for allowing the use of the computer cluster at the Rowett Research Institute for the random forests analyses.

Grants

This work was supported by the Centre for Human Nutrigenomics (A.G. Heidema), TI Food and Nutrition (W. Rodenburg), and the ministry of Agriculture, Food Quality and Nature Management of The Netherlands. J. Keijer is a member of Mitofood.

References

1. Afman L, Muller M: Nutrigenomics: from molecular nutrition to prevention of disease. J Am Diet Assoc 2006, 106:569–576.
2. de Boer VC, van Schothorst EM, Dihal AA, van der Woude H, Arts IC, Rietjens IM, Hollman PC, Keijer J: Chronic quercetin exposure affects fatty acid catabolism in rat lung. Cell Mol Life Sci 2006, 63:2847–2858.

3. Patsouris D, Reddy JK, Muller M, Kersten S: Peroxisome proliferatoractivated receptor alpha mediates the effects of high-fat diet on hepatic gene expression. Endocrinology 2006, 147:1508–1516.
4. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003, 34:267–273.
5. Norris AW, Kahn CR: Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates. Proc Natl Acad Sci USA 2006, 103:649–653.
6. Segal E, Friedman N, Kaminski N, Regev A, Koller D: From signatures to models: understanding cancer using microarrays. Nat Genet 2005, 37, Suppl:S38–S45.
7. Breitling R, Armengaud P, Amtmann A, Herzyk P: Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 2004, 573:83–92.
8. Yoon S, Yang Y, Choi J, Seong J: Large scale data mining approach for gene-specific standardization of microarray gene expression data. Bioinformatics 2006, 22:2898–2904.
9. Allison DB, Cui X, Page GP, Sabripour M: Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006, 7:55–65.
10. Verducci JS, Melfi VF, Lin S, Wang Z, Roy S, Sen CK: Microarray analysis of gene expression: considerations in data mining and statistical treatment. Physiol Genomics 2006, 25:355–363.
11. Brazhnik P, de la Fuente A, Mendes P: Gene networks: how to put the function in genomics. Trends Biotechnol 2002, 20:467–472.
12. Slonim DK: From patterns to pathways: gene expression data analysis comes of age. Nat Genet 2002, 32, Suppl:502–508.
13. Barry WT, Nobel AB, Wright FA: Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005, 21:1943–1949.
14. Chen JJ, Wang SJ, Tsai CA, Lin CJ: Selection of differentially expressed genes in microarray data analysis. Pharmacogenomics J 2006, 7:212–220.
15. Hochberg Y, Benjamini Y: More powerful procedures for multiple significance testing. Stat Med 1990, 9:811–818.
16. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001, 98:5116–5121.
17. Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004, 32:D258–D261.
18. Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000, 28:27–30.
19. Khatri P, Draghici S: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005, 21:3587–3595.
20. Draghici S, Sellamuthu S, Khatri P: Babel's tower revisited: a universal resource for cross-referencing across annotation databases. Bioinformatics 2006, 22:2934–2939.
21. Quackenbush J: Computational analysis of microarray data. Nat Rev Genet 2001, 2:418–427.
22. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999, 96:2907–2912.
23. Valafar F: Pattern recognition techniques in microarray data analysis: a survey. Ann NY Acad Sci 2002, 980:41–64.
24. Werner T: Regulatory networks: linking microarray data to systems biology. Mech Ageing Dev 2007, 128:168–172.

25. Lee HK, Braynen W, Keshav K, Pavlidis P: ErmineJ: tool for functional analysis of gene expression data sets. BMC Bioinformatics 2005, 6:269.
26. Tomfohr J, Lu J, Kepler TB: Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics 2005, 6:225.
27. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T: Pathway mapping tools for analysis of high content data. Methods Mol Biol 2007, 356:319–350.
28. Rubin E: Circumventing the cut-off for enrichment analysis. Brief Bioinform 2006, 7:202–203.
29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005, 102:15545–15550.
30. Majumder PK, Febbo PG, Bikoff R, Berger R, Xue Q, McMahon LM, Manola J, Brugarolas J, McDonnell TJ, Golub TR, Loda M, Lane HA, Sellers WR: mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways. Nat Med 2004, 10:594–601.
31. Breiman L: Random Forest. Machine Learning 2001, 45:5–32.
32. Huang X, Pan W, Grindle S, Han X, Chen Y, Park SJ, Miller LW, Hall J: A comparative study of discriminating human heart failure etiology using gene expression profiles. BMC Bioinformatics 2005, 6:205.
33. Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S: Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol 2005, 18:547–557.
34. Diaz-Uriarte R, Alvarez de Andres S: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006, 7:3.
35. Lee JW, Lee JB, Park M, Song SH: An extensive comparison of recent classification tools applied to microarray data. Comput Stat Data Anal 2005, 48:869–885.
36. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004, 5:32.
37. Smyth GK, Yang YH, Speed T: Statistical issues in cDNA microarray data analysis. Methods Mol Biol 2003, 224:111–136.
38. Pellis L, Franssen-van Hal NL, Burema J, Keijer J: The intraclass correlation coefficient applied for evaluation of data correction, labeling methods, and rectal biopsy sampling in DNA microarray experiments. Physiol Genomics 2003, 16:99–106.
39. Liaw A, Wiener M: Classification and regression by randomForest. R News 2002, 2:18–22.
40. Breiman L: Fortran Code for Random Forests. [www.stat.berkeley.edu/user/breiman/randomforests/].
41. Cox DR, Hinkley DV: Theoretical Statistics. London: Chapman and Hall; 1974.
42. Lyons-Weiler J, Pelikan R, Zeh HJ, Whitcomb DC, Malehorn DE, Bigbee WL, Hauskrecht M: Assessing the statistical significance of the achieved classification error of classifiers constructed using serum peptide profiles, and a prescription for random sampling repeated studies for massive high-throughput genomic and proteomics studies. Cancer Informatics 2005, 1:53–77.
43. GeneSrF. Gene selection with random forests. CNIO Bioinformatics Unit. [http://genesrf.bioinfo.cnio.es].
44. CRAN. [http://cran.r-project.org/].
45. R Development Core Team: R: a language and environment for statistical computing. [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna, 2004.
46. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007, 8:25.

47. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B: Microarray data mining with visual programming. Bioinformatics 2005, 21:396–398.
48. Dopazo J: Functional interpretation of microarray experiments. OMICS 2006, 10:398–410.
49. Goeman JJ, Buhlmann P: Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007, 23:980–987.
50. Parast MM, Otey CA: Characterization of palladin, a novel protein localized to stress fibers and cell adhesions. J Cell Biol 2000, 150:643–656.

Chapter 6

Developing a discrimination rule between breast cancer patients and controls using proteomics mass spectrometric data: a three-step approach

A Geert Heidema[1,2,3]
Nico Nagelkerke[4,5]

[1] Department of Human Biology, Maastricht University, Maastricht, The Netherlands;
[2] Centre for Nutrition and Health, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands;
[3] Division of Human Nutrition, Wageningen University and Research Centre, Wageningen, The Netherlands;
[4] Department of Community Medicine, United Arab Emirates University, Al Ain, UAE;
[5] Department of Medical Microbiology, University of Manitoba, Winnipeg, Canada.

Statistical Applications in Genetics and Molecular Biology 2008, 7(2):5.

Abstract

To discriminate between breast cancer patients and controls, we used a three-step approach to obtain our decision rule. First, we ranked the mass/charge values using random forests, because it generates importance indices that take possible interactions into account. We observed that the top ranked variables consisted of highly correlated contiguous mass/charge values, which were grouped in the second step into new variables. Finally, these newly created variables were used as predictors to find a suitable discrimination rule. In this last step, we compared three different methods, namely Classification and Regression Tree (CART), logistic regression and penalized logistic regression. Logistic regression and penalized logistic regression performed equally well and both had a higher classification accuracy than CART. The model obtained with penalized logistic regression was chosen as we hypothesized that this model would provide a better classification accuracy in the validation set. The solution had a good performance on the training set with a classification accuracy of 86.3%, and a sensitivity and specificity of 86.8% and 85.7%, respectively.

Introduction

To develop a decision rule that discriminates well between individuals with breast cancer (cases) and individuals without (controls), we applied a three-step approach, viz. 1) variable selection/reduction; 2) synthesis of new variables by grouping selected variables to make use of the correlation structure; and 3) actual discrimination and classification on the basis of the variables developed in step 2. In step 1, in order to make a reduction in the large numbers of variables present in the dataset we applied random forests (RF). RF appears to be the most appropriate method for prioritizing variables and selection of a small set of most important variables, i.e. variables that appear to hold promise of having discriminatory power *in conjunction with other variables*. The latter clause is important, as selection of variables on the basis of individual discriminatory power is unsatisfactory in view of the correlation between many variables. Also, as it is based on cross-validation, one can expect the selected variables to work (i.e. to be discriminatory) not only in the training dataset, but also from validation datasets collected from different patients and controls from the same population. RF was developed by Breiman [1]. This machine learning approach has proven to have excellent performance in many classification tasks, and is now available as an off-the-shelf method. RF has shown to outperform other classification methods in applications to microarray data [2] and mass spectrometry data [3]. One of the features of RF is that it provides a measure of importance for each of the variables, referred to as the importance index. The importance index was used to prioritize and select the variables that best discriminate between cases and controls.

Contiguous variables with approximately identical mass are highly correlated due to physical properties and the smoothing applied in the pre-processing steps. Therefore, it can be expected that among the highly prioritized variables highly correlated, contiguous variables will be present. Therefore, in the second step we searched over the most important variables whether groups of highly correlated variables would be present. These highly correlated variables can then be grouped into a new variable. In the third step we used these newly created variables as predictors and applied different methods to find a suitable discrimination rule. The methods compared at this step are Classification and Regression Tree (CART) [4], logistic regression and penalized logistic regression [5, 6]. The decision rules obtained by the different methods and their classification performance were compared and the decision rule with the best performance was finally chosen to be applied to the validation set.

Methods

Step 1: Prioritization of variables by random forests

To reduce the number of variables to be used to make a decision rule, we used RF to prioritize and select the apparently best discriminating candidate variables in the first step.

In RF, an ensemble of tree models is used to predict case-control status (bagging). Each tree recursively splits the total dataset into smaller and more homogeneous subgroups of cases and controls, whereby the total sample for each tree is obtained by bootstrap sampling. With bootstrap sampling, sampling is performed with replacement and some individuals are sampled more than once while others are left out, while keeping the

bootstrap sample size the same as that of the original sample. This method involves cross-validation as the (bootstrap) sampled observations are used to construct the classification tree whereas a prediction is obtained for each left-out individual. Aggregating the predictions over the different trees in which the individual was left-out, a prediction for this individual is obtained for the ensemble of trees, which is called the forest. The proportion of misclassified cases and controls provides the prediction error of the forest. Another important feature is that the predictor that gives the best partitioning in cases and controls at a certain split is not selected from the total number of predictors but from a smaller random sample of predictors. This parameter is referred to as m_{try}. We used the default value for m_{try}, which is the square root of the number of variables to be analyzed in the dataset (in this dataset equal to 105). Multiple thousands of trees in the forest are needed to obtain stable estimates of the importance indices [7]. Also, each tree captures only the possible interactions for the variables selected by that tree only, and large numbers of trees are required to capture as many interactions as possible. Therefore, the number of trees in the forest was set to 30,000 for each of the different analyses. We performed several analyses with RF to verify whether the ranking of the variables by their importance index did not change over the different analyses. This is done by using different seed values for the different analyses (the seed value controls the random number generator).

RF provides an importance index for each variable by comparing the predictive performance of the forest for all variables with the predictive performance of the forest for all variables but with the values for one variable randomly permuted for the left-out individuals. Larger differences in the predictive performance indicate more important variables. Permuting the predictor values for the left-out individuals does not only remove the association between the permuted predictor and the outcome variable, but also the interaction effects of the permuted predictor with other predictors, if present. Thereby, the interactions of the predictor with other predictors are taken into account in the importance index. We used the importance index as a first step to prioritize and select the best discriminating variables.

For the RF analyses we used the R-package randomForest written by Liaw and Wiener [8, 9], freely available from the CRAN website (http://cran.r-project.org/). Because the predictors are all of the same type, the RF variable importance indices obtained with the randomForest R-package can be used [10]. This R-package is based on the original FORTRAN code from Breiman et al. [11] (freely available at www.stat.berkeley.edu/users/breiman/randomforests/).

Step 2: Grouping highly correlated variables

Among the highest prioritized variables we tried to exploit the correlation between variables. Adjacent variables were very highly correlated (generally $r>0.9$) which made adjacent variables almost duplicate measurement of the same underlying “parameter”. If groups of adjacent variables were identified, we combined these variables into a new variable by taking the sum of the variables. Small “gaps” were ignored, i.e. variables of which both its neighbours were selected, were included even if that variable itself was not selected.

Step 3: Obtaining the decision rule

The newly created variables, each consisting of a group (sum) of highly correlated variables, were used as candidate predictors for case-control status. We tried the following methods to predict the individuals in the calibration dataset: CART, logistic regression and penalized logistic regression. For CART we used the program QUEST [12], which is freely available at http://www.stat.wisc.edu/~loh/quest.html. To perform logistic regression analysis, SPSS version 13.0 was used (SPSS, Inc., Chicago, Illinois). For penalized logistic regression, we applied the R-package brlr written by Firth [6]. This method penalizes the likelihood by the Jeffrey's prior, and has the effect of "mildly" shrinking parameter estimates to 0. The decision rule obtained by the method with the best classification performance was chosen to obtain the prediction for the individuals in the validation dataset.

Results

Step 1: Prioritization of variables by random forests

The prioritization of variables by RF for two different seed values are shown in figure 1. For both seed values the same variables were highly prioritized. Thus RF provides similar results over different analyses, indicating the robustness of the method.

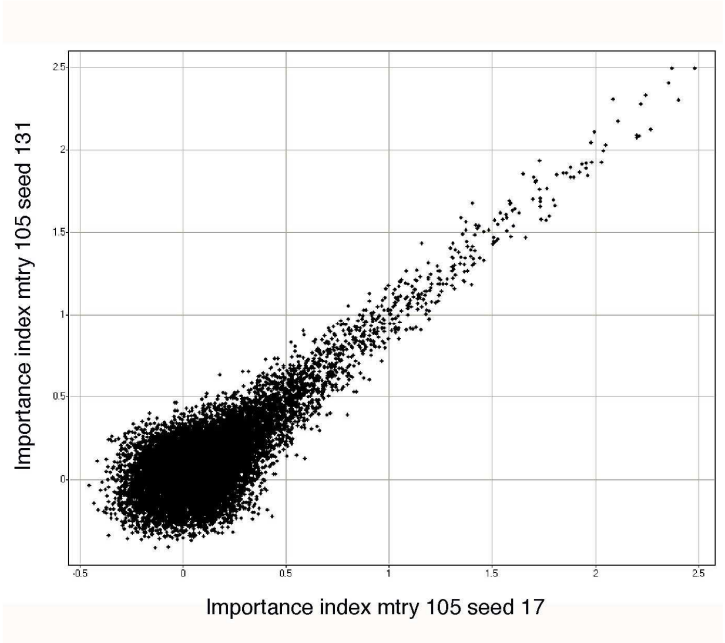


Figure 1: RF results. Prioritization of m/z values by their importance index for two different seed values.

Step 2: Grouping highly correlated variables

As expected, visual inspection of the most important variables showed that highly prioritized m/z values consisted of different groups of contiguous variables. Therefore, we combined adjacent variables into a new variable, summing the scores of the individual variables. In this way, nine new variables were formed (see Box 1).

Box 1: New variables (Y1-Y9) formed by summing the scores of adjacent individual variables. The numbers of the individual variables represent mass/charge values.

Y1 = v3454 + v3455 + v3456 + v3457 + v3458 + v3459 + v3460.

Y2 = v3496 + v3497 + v3498 + v3499.

Y3 = v3830 + v3831 + v3832 + v3833 + v3834 + v3835 + v3836.

Y4 = v3844 + v3845 + v3846 + v3847 + v3848 + v3849 + v3850 + v3851 + v3852 + v3853 + v3854 + v3855 + v3856 + v3857 + v3858.

Y5 = v3924 + v3925 + v3926 + v3927 + v3928 + v3929 + v3930.

Y6 = v6607 + v6608 + v6609 + v6610.

Y7 = v9531 + v9532 + v9533 + v9534 + v9535.

Y8 = v5380 + v5381.

Y9 = v6606 + v6607.

Step 3: Obtaining the decision rule

The nine variables obtained in step 2 were used as predictors to obtain a decision rule. At this step we compared the classification performance of CART, logistic regression and penalized logistic regression. Table 1 shows the classification accuracy for the different methods.

Table 1: Classification accuracy obtained with CART, logistic regression and penalized logistic regression.

Method	Classification accuracy (%)
CART	78.1
Logistic regression	86.3
Penalized logistic regression	86.3

Logistic regression and penalized logistic regression both had a higher classification accuracy compared to CART. For both logistic regression and penalized logistic regression neither interactions nor logarithmically transformed variables did improve the classification performance and were therefore not included in the final model. Both types of logistic regression performed equally well, in fact they gave identical classification results and we chose the model obtained with penalized logistic regression as we hypothesized that this solution, as its estimators have been designed to have less bias, would give a better

classification accuracy for the validation set. Thus, using penalized logistic regression, we obtained the following solution:

Logit{Pr(individual belongs to group (Cases))} = 1572.13 – 57.36·Y1 + 4.98·Y2 – 13.32·Y3 – 13.04·Y4 + 4.36·Y5 + 69.07·Y6 – 19.88·Y7 + 124.08·Y8 – 1191.26·Y9

Performance on training data

The cross-classification table is shown in table 2. The classification accuracy equals 86.3%. For sensitivity and specificity, very similar percentages were obtained (86.8% and 85.7%, respectively).

Table 2: Cross-classification table, based on the decision rule obtained with penalized logistic regression at the third step of the three-step approach.

	True group control	True group cases
Assigned group control	66	10
Assigned group cases	11	66

Discussion

The three-step approach we applied to obtain a decision rule to discriminate between cases of breast cancer and controls has several advantages. The use of RF in the first step has the advantage that the interdependence between variables is taken into account in the importance index, and therefore in the prioritization and selection of variables. Also, standard stepwise procedures tend to reject or select variables on the basis of their individual discriminating power, which may be far from optimal in a context of many highly correlated variables. The high correlation between (prioritized) variables is also taken into account in the second step by combining adjacent variables into new variables, with the idea that these variables essentially measure the same "peak" or other feature and that therefore the measurement errors of these new "sum" variables is less than that of individual variables. Furthermore, a clear interpretation of the predictors on which the decision rule is based can be made in the third step.

However, there are also limitations to this three-step approach. There is a methodological discrepancy or disconnect between using trees in variable selection and using logistic regression in the final (third step) analysis. As logistic regression was chosen because it clearly outperformed CART, one may conjecture whether in the selection phase a logistic regression approach, but one that would make use of bootstrapping and cross-validation in a similar way as RF would have been better at selecting variables. Unfortunately, this method is not available off-the-shelf using readily available software. Also, the grouping of variables in the second step of our analysis was done, more or less *ad hoc*, by eye and hand, and therefore this step is not amenable to cross-validation. This step should have been formalized and automated and used in cross-validation. Finally, Jeffreys prior may well be too "flat" and heavier shrinkage might have yielded classifiers with great predictive efficiency.

Another limitation of our approach, or any other approach that is based on variable selection, is that it is based on an untested assumption, *viz.* that the classification problem is "sparse" in the sense that only a small minority of variables have any discriminating power and that the rest is essentially "noise". While this seems to be supported by the finding that only the importance index of variables with a high importance index tends to be reproducible across different runs of the random forests program, this is by no means certain. If this "sparseness" does not hold true then the additional discriminating power of many weakly informative variables is ignored by our approach. With such a limited sample size however, the task of making effective use of such variables would seem daunting.

A further limitation of our three step approach is lack of methodological coherence. This was largely due to our objective of developing an easy-to-apply discrimination score, and our idea that we had to take into account the correlation structure among (neighbouring) variables. However, this mixture of methods makes it harder to identify the causes of misclassifications. These are much easier to identify and perhaps correct when only a single method is used, for example RF. For the application of only RF the selection of variables for classification as performed by Diaz et al. [2] could be used. Although this approach leads to a small set of variables that still has good performance, it does not lead to readily usable classification rules for clinical diagnostic purposes, and neither does it give rise to classification rules that are easy to interpret.

Finally we want to address the possible sensitivity of our approach to experimental effects. Plates on which the experiments were run were not known to us. Thus our method may be sensitive to "plate effects". If some plates yield (locally) systematically higher or lower values than other plates this may influence and bias the classification results. Perhaps, instead of new synthetic variables consisting of sums of variables, sums of contrasts, i.e. sums of signed variables, with as many positive as negative signs, would be less sensitive to such plate effects. Preferably, perhaps, any variables with negative signs should be matched with, and chosen relatively close to variables with a positive sign. However, of course, such control variables should be chosen sufficiently distant to avoid strong correlations with their positive "matches".

References

1. Breiman L: Random forests. Machine learning 2001, 45:5-32.
2. Díaz-Uriarte R, Alvarez de Andrés S: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006, 7:3.
3. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H: Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003, 19:1636-1643.
4. Breiman L, Friedman JH, Olshen RA, Stone CJ: Classication and RegressionTrees, 1984, Wadsworth, Belmont.
5. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic-regression. Appl Stat J of the Royal Stat Soc Series C 1992, 41(1):191-201.
6. Firth D: Bias reduction of maximum likelihood estimates. Biometrika 1993, 80:27–38.
7. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004, 5:32.
8. Liaw A, Wiener M. 2002. Classification and regression by random-Forest. Rnews 2:18-22.

9. R Development Core Team: R: a language and environment for statistical computing. [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna, 2004.
10. Strobl C, Boulesteiz A-L, Zeileis A, Hothorn T: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 2007, 8:25.
11. Breiman L, Cutler A: Random forests. Version 4.0. 2003. [http://www.stat.berkeley.edu/users/breiman/RandomForests/].
12. Loh W-Y, Shih Y-S: Split selection methods for classification trees. Statistica Sinica 1997, 7:815-840.

Chapter 7

The association of 83 plasma proteins with CHD mortality, BMI, HDL- and total cholesterol in men: applying multivariate statistics to identify proteins with prognostic value and biological relevance

A Geert Heidema[1,2,3]
Uwe Thissen[4,5]
Jolanda MA Boer[2]
Freek G Bouwman[1]
Edith JM Feskens[3]
Edwin CM Mariman[1]

[1] Department of Human Biology, Maastricht University, Maastricht, The Netherlands;
[2] Centre for Nutrition and Health, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands;
[3] Division of Human Nutrition, Wageningen University and Research Centre, Wageningen, The Netherlands;
[4] TNO Quality of Life, Zeist, The Netherlands;
[5] Dutch nutrigenomics consortium of the Top Institute Food and Nutrition (TIFN), Wageningen, The Netherlands.

Submitted.

Abstract

In this study we applied the multivariate statistical tool Partial Least Squares (PLS) to analyze 83 plasma proteins in relation to coronary heart disease (CHD) mortality and the intermediate endpoints body mass index, HDL-cholesterol and total cholesterol. From a Dutch monitoring project for cardiovascular disease risk factors men who died of CHD between initial participation (1987-1991) and end of follow up (January 1, 2000) (N=44) and matched controls (N=44) were selected. Baseline plasma concentrations of proteins were measured by a multiplex immunoassay. Applying PLS we identified 15 proteins with prognostic value for CHD mortality and sets of proteins associated with the intermediate endpoints. Subsequently, sets of proteins and intermediate endpoints were analyzed together by Principal Components Analysis, indicating that proteins involved in inflammation explained most of the variance, followed by proteins involved in metabolism and proteins associated with total cholesterol. This study is one of the first in which the association of a large number of plasma proteins with CHD mortality and intermediate endpoints is investigated by applying multivariate statistics, providing insight in the relationships among proteins, intermediate endpoints and CHD mortality, and a set of proteins with prognostic value.

Introduction

Coronary heart disease (CHD) often manifests at a later stage of life. CHD is mainly caused by narrowed arteries due to atherosclerosis, diminishing the supply of blood, oxygen and nutrients to the heart [1], which finally can lead to a myocardial infarction. Besides treatment of CHD, detecting the risk of CHD at a preclinical stage is of vital importance. To realize this objective, proteins that have prognostic value for CHD need to be identified. Furthermore, insight in the relationships between plasma proteins and important risk factors involved in the etiology of CHD (e.g. high body mass index (BMI) [2], low HDL-cholesterol (HDL-C) [3] and high total cholesterol (total-C) [1]) can provide useful information for prevention of CHD later in life.

Nowadays, technological advances in proteomics such as multiplex assays provide the opportunity to simultaneously measure large numbers of protein concentrations in plasma. This enables researchers to study, besides individual associations, the relationships of groups of proteins with the outcome of interest. To analyze proteomic data Partial Least Squares (PLS) is a suitable multivariate statistical tool [4]; in contrast to univariate statistics, PLS takes the information of all proteins into account with respect to a certain endpoint. In this study we applied PLS to investigate in men the association of 83 proteins with CHD mortality, and the intermediate endpoints BMI, HDL-C and total-C, with the objectives to identify a set of proteins with prognostic value for CHD mortality and to identify sets of proteins associated with BMI, HDL-C and total-C. Subsequently we applied Principal Component Analyses (PCA) [5] to interpret the relationships between all identified proteins, intermediate endpoints and CHD mortality.

Methods

Study population

For this study 44 male subjects who died of CHD and 44 male controls were selected from a monitoring project for cardiovascular disease risk factors [6], which was carried out in three towns (Amsterdam, Doetinchem and Maastricht) in the Netherlands. Cases were randomly selected from the 162 men who died of CHD (ICD-9 410-414 or ICD-10 I20-I25) between their baseline examination (1987-1991) and end of follow up (January 1, 2000). On average the cases died 5.83 ± 3.05 years after baseline examination. From the men who did not experience a CHD event during follow up controls were randomly selected, but were matched for town of investigation, age, smoking status and year of baseline examination. Informed consent for using the stored blood samples for research purposes was given by all subjects. Characteristics of the study population are shown in table 1.

Table 1: Characteristics of the study population.

	CHD mortality		p-value
	Cases (N=44)	Controls (N=44)	
Age (years)	51.9 ± 5.37	51.5 ± 5.86	
BMI (kg/m^2)	27.1 ± 3.58	25.8 ± 3.41	0.10
HDL-C (mmol/l)	0.97 ± 0.23	1.13 ± 0.31	0.01**
Total-C (mmol/l)	6.41 ± 1.27	5.94 ± 1.04	0.07

For cases and controls their mean ± sd are shown
** Significant at the 0.01 level

Measurements

Intermediate endpoints

Weight at baseline was measured at the Municipal Health Centre, whereby subjects were wearing light indoor clothing without shoes. To obtain the BMI at baseline, weight at baseline was divided by height squared (in kg/m^2).

At baseline non-fasting blood samples were obtained in EDTA-coated vacutainer tubes. Plasma total-C and HDL-C were determined enzymatically using a Boehringer test-kit within three weeks after storage [7]. HDL-C was determined after precipitation of ApoB containing lipoproteins with magnesium phosphotungstate [8].

Proteins

The concentrations of 89 proteins were measured by Rules-Based Medicine (RBM) in non-fasting plasma by a multiplex immunoassay (HumanMAP Version 1.6, Rules-Based Medicine, Inc., Austin, TX). The set of proteins present on this assay consists of selected factors that are implicated in different types of diseases.

Prior to the statistical analyses, 7 proteins were removed from the dataset for which more than half of the samples were not measurable on the standard curve. For other proteins, values not measurable on the standard curve were imputed with 0.1*Least Detectable Dose. For 62 proteins all samples were measurable, whereas for 13, 4 and 3 proteins the number of samples not measurable was between 1 and 5, between 6 and 10, or more than 10, respectively. ApoB levels were additionally measured on a Hitachi 912 autoanalyser (Roche, Lelystad, The Netherlands) using a commercially available kit (Roche cat. nr. 1551779). Together, 83 proteins were included in the statistical analyses.

Statistical analyses

Univariate analyses

The difference in mean plasma levels between cases and controls for each protein was tested with the conventional t-test. If proteins were not normally distributed (determined by Kolmogorov-Smirnov test), the Mann-Whitney U-test was applied. BMI, HDL-C and total-C were slightly skewed for the total population, therefore Spearman correlation coefficients were calculated between proteins concentrations and the values of these intermediate

endpoints, and for the associations among the intermediate endpoints. As the intermediate endpoints were normally distributed for cases and controls separately, differences between the values of the intermediate endpoints for CHD cases and controls were determined by t-tests. These analyses were performed using SAS software version 9.1 (SAS institute Inc., Cary, NC, USA).

PLS-DA and PLS-regression

Partial least squares (PLS) is a multivariate method aimed at relating measured data to an outcome of interest. PLS is able to handle large numbers of variables in moderate to small sample sizes. It reduces the dimensionality of the data by constructing latent components, in such a way that these components have maximal covariance with the outcome variable whereas the latent components themselves are uncorrelated. PLS can be applied to classify categorical outcome variables using PLS in combination with discriminant analyses (DA), referred to as PLS-DA, and in regression problems to analyze continuous outcome variables, referred to as PLS-regression. More detailed information on PLS-DA and PLS-regression can be obtained from Boulesteix and Strimmer [4]. In this study we applied PLS-DA to analyze the relation of 83 proteins with CHD mortality, whereas PLS-regression was applied to analyze the relation of the 83 proteins with baseline levels of BMI, HDL-C and total-C.

To take the differences in measurement scales of the proteins into account, proteins were auto-scaled before performing the PLS analyses, resulting in a mean of 0 and a standard deviation of 1 for each protein [9]. Double cross-validation was applied to unbiasedly determine both the number of latent components in the PLS-model and the classification error (PLS-DA) or R^2 (PLS-regression) of the model [10, 11]. Briefly, first 10-fold cross-validation was applied, splitting the total dataset into ten subsets. A 9-fold cross-validation was applied for each 9/10 of the ten subsets to determine the number of latent components for the PLS-model that provided the smallest error (or largest R^2). The remaining 1/10 of the data of the 10-fold cross-validation was used as a test set to evaluate the error of the PLS-model. The 10 errors obtained for each of the PLS-models were averaged to obtain an unbiased estimate of the classification error (PLS-DA) or R^2 (PLS-regression). A permutation test [12, 13] using 1000 permuted datasets was applied to obtain the statistical significance for the PLS-model of CHD.

Proteins were considered to be associated with the different endpoints if the confidence interval (mean ± 2*standard error) of their model coefficient deviated from 0. This cut-off corresponds to a Relative Standard Deviation (RSD: standard deviation divided by the mean) smaller than 0.5. The standard deviations of the model coefficients have been determined as described by Faber [14].

PLS-DA and PLS-regression analyses were performed in Matlab 7.3.0 (The Mathworks, Natick, MA, USA) and the PLS Toolbox 3.5.4 (Eigenvector Research, Manson, WA, USA).

PCA

Principal components analyses (PCA) [5] was applied to visualize the relationships among all selected proteins by PLS-DA and PLS-regression together with the intermediate endpoints. PCA is based on different statistical concepts than PLS, because these

techniques serve different aims: whereas PLS constructs latent components that have maximal covariance with the outcome variable, PCA constructs factors that maximally explain the variance in the data, without relating the factors to another outcome variable (or set of outcome variables). We applied GeneMaths XT version 1.6.1 (Applied Math) software to obtain the PCA plots.

Results

Associations among endpoints

CHD cases had on average 0.15 mmol/l lower HDL-C levels compared to controls (95% confidence interval: -0.27– -0.04, p=0.01, see table 1). They also tended to have a higher BMI (mean difference [95% confidence interval]=1.3, [-0.23–2.73], p=0.10) and higher total-C levels (mean difference [95% confidence interval]=0.46, [-0.03–0.95], p=0.07). A significant inverse relationship was also present between BMI and HDL-C (r=-0.41, p<0.0001), whereas no significant correlations were observed of BMI and HDL-C with total-C (r=0.05, p=0.61 and r=-0.11, p=0.29, respectively).

Univariate associations of proteins with CHD and intermediate endpoints

The univariate relationships between the 83 proteins and CHD mortality, BMI, HDL-C and total-C are shown in the supplemental tables 1 and 2 (see supplemental data). Eight proteins were significantly (p<0.05) related to CHD mortality (ApoA1, ApoB, β-2 Microglobulin, CRP, PAI-1, PAPP-A, VCAM-1 and VEGF). Except for ApoA1, higher concentrations were observed for these proteins in CHD cases compared to controls. For BMI, HDL-C and total-C the number of proteins found to be significantly (p<0.05) associated was equal to 11, 22 and 7, respectively. As BMI and HDL-C are inversely related, proteins that were significantly associated with both endpoints showed opposite relationships. The proteins most (p<0.005) related to BMI included complement 3, ferritin, IL-1ra, insulin, leptin (all positively), growth hormone and SHBG (both negatively). Adiponectin and ApoA1 were most (p<0.005) positively associated with HDL-C, whereas ApoB, complement 3, CRP, IL-1ra, insulin, PAI-1, and TBG were most negatively associated. Finally, total-C was highly correlated (r= 0.87) with ApoB, and to a lesser extent positively correlated with MCP-1 (r= 0.41).

PLS-DA and PLS-regression

Reducing the noise of non-important proteins

The PLS-model for CHD including all proteins was not statistically significant (classification accuracy= 53%, p= 0.63). However, proteins in the model (see figure 1A) for which the 95%-confidence interval of their model coefficient deviated from 0 were previously found to be biologically relevant in relation to CHD in the literature, for example ApoA1 [15], CRP [16], VCAM1 [17] and VEGF [18, 19]. This may indicate that the noise of non-important proteins obscures the signals of important proteins. The PLS-models including all proteins also had limited predictive performance for BMI, HDL-C and

total-C (R^2=0.36, R^2=0.14 and R^2=0.30, respectively; see figures 1B-D). As for CHD, biologically relevant proteins for these intermediate endpoints were also found. For example, leptin was found to be strongly positively correlated with BMI, ApoA1 with HDL-C and ApoB with total-C.

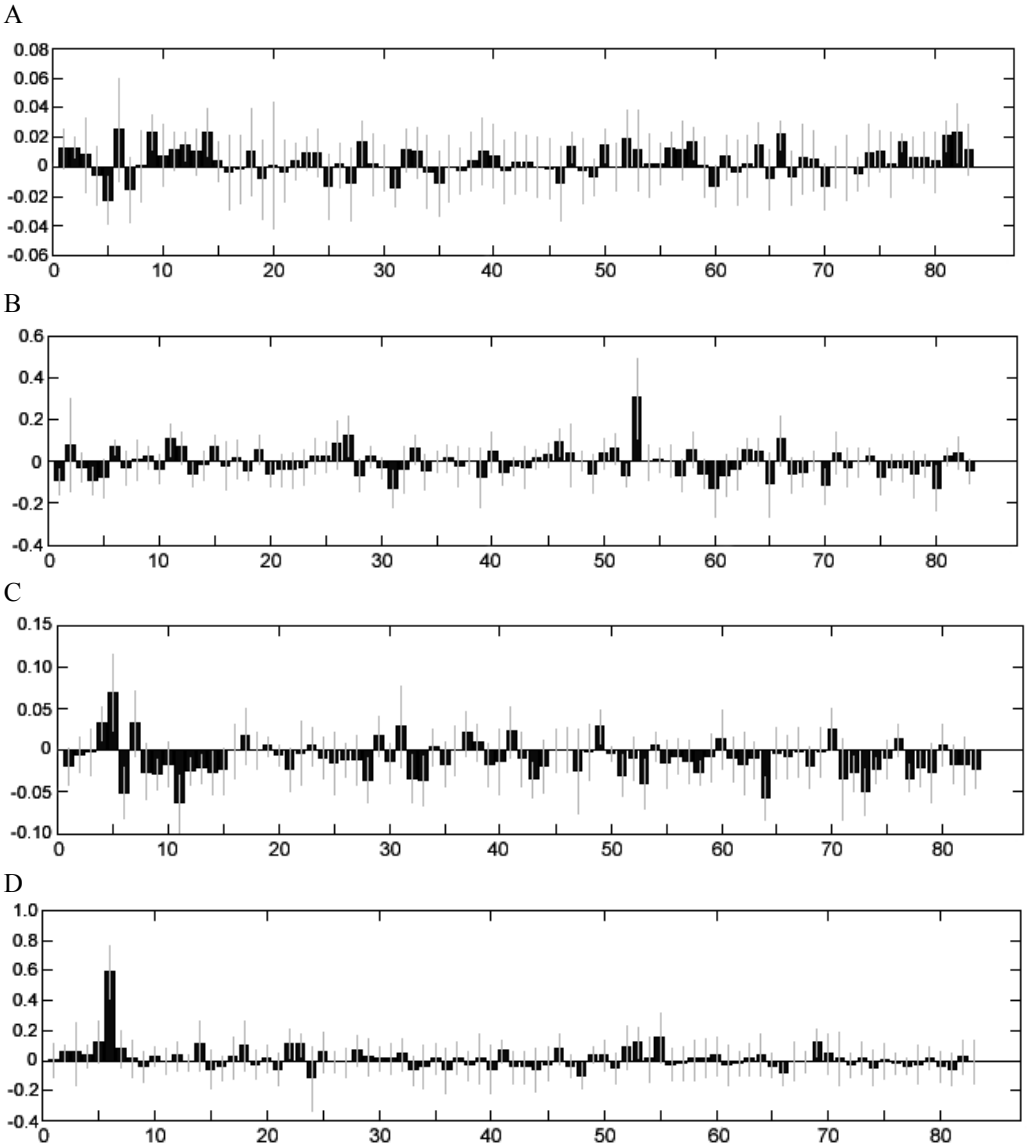
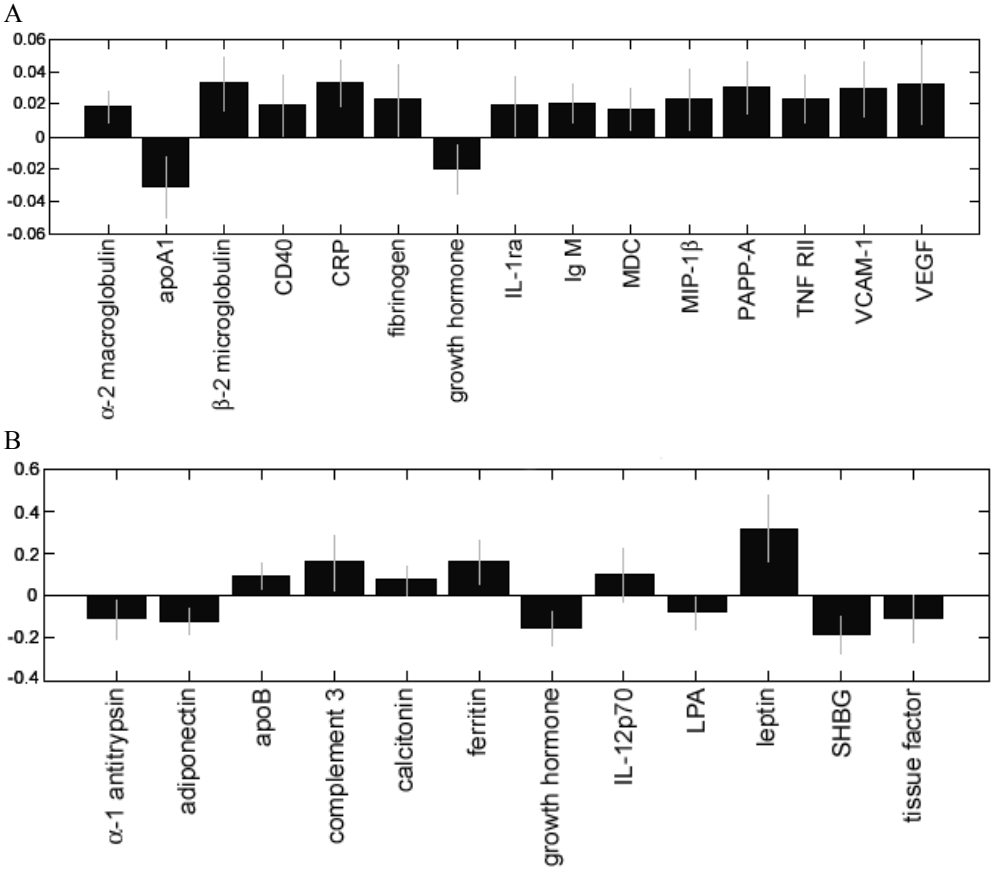


Figure 1: PLS model of 83 proteins in relation to CHD mortality (A), BMI (B), HDL-C (C) and total-C (D). The regression vector ± 2*standard deviation is shown. Numbers correspond to the numbers of the proteins listed in the supplemental table (see supplemental data).

As these results are supported by results previously reported in the literature [20-22], we reanalyzed the relationship between the proteins and the different endpoints, reducing the noise by taking only those proteins with a relative standard deviation (RSD) of smaller than 0.5 into account. For total-C no PLS-model could be obtained applying a cut-off of RSD<0.5, therefore for total-C a cut-off of RSD<1.0 was used. Within the reduced PLS models, proteins were again considered to be important if their confidence interval (mean ± 2*standard error) within the model deviated from 0.

Association of proteins with CHD mortality

The newly obtained PLS model for CHD including only those proteins with RSD<0.5 improved and was statistically significant (classification accuracy= 65%, p=0.038). In this reduced multivariate model, 15 proteins were associated with CHD mortality, with inverse relationships for ApoA1 and growth hormone and positive relationships for other proteins (see figure 2A and table 2).



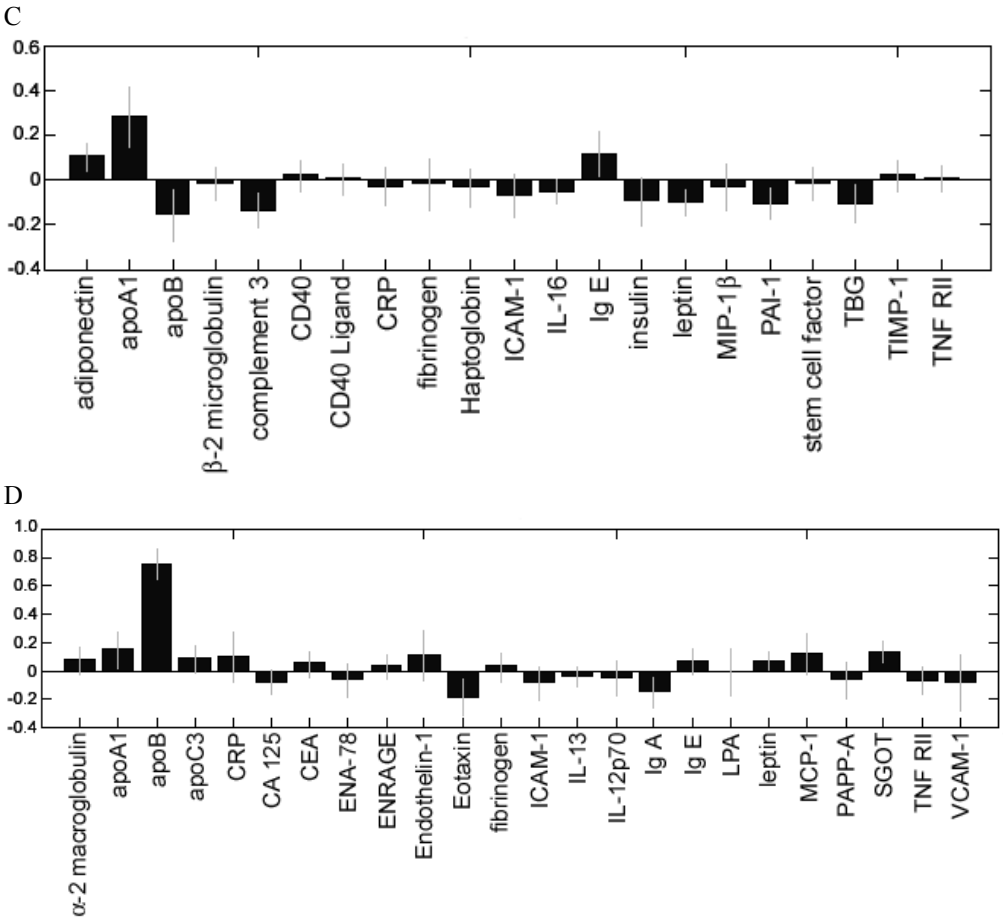


Figure 2: PLS model of proteins in relation to CHD mortality (A), BMI (B), HDL-C (C) using cut-off point of RSD<0.5 and total-C (D) using cut-off point of RSD<1.0. The regression vector ± 2*standard deviation is shown.

Six of these proteins (ApoA1, β-2 Microglobulin, CRP, PAPP-A, VCAM-1, VEGF) also showed an association with CHD in the univariate analysis, while the other proteins univariately only showed a trend (IL-1ra, TNF RII) or were found to be non-significant (α-2 macroglobulin, CD40, fibrinogen, growth hormone, Ig M, MDC, MIP-1β).

Table 2: PLS-models for the different endpoints including proteins selected based on cut-off of RSD<0.5 to reduce noise, except for the PLS-model of total-C, which is obtained with RSD<1.0.

CHD mortality			BMI			HDL-C			Total-C		
Protein	sign	univariate p-value	protein	sign	univariate p-value	Protein	sign	univariate p-value	protein	sign	univariate p-value
α-2 Macroglobulin	+	0.61	α-1 Antitrypsin	–	0.11	Adiponectin	+	0.0007	α-2 Macroglobulin		0.68
ApoA1	–	0.02	Adiponectin	–	0.01	ApoA1	+	<0.0001	ApoA1	+	0.38
β-2 Microglobulin	+	0.009	ApoB	+	0.10	ApoB	–	0.0004	ApoB	+	<0.0001
CD40	+	0.24	Complement 3	+	0.0001	B-2 Microglobulin		0.005	ApoC3		0.46
CRP	+	0.01	Calcitonin	+	0.15	Complement 3	–	<0.0001	CRP		0.13
Fibrinogen	+	0.12	Ferritin	+	0.002	CD40		0.04	CA 125		0.39
Growth Hormone	–	0.22	Growth Hormone	–	0.003	CD40 Ligand		0.35	CEA		0.37
IL-1ra	+	0.10	IL-12p70		0.13	CRP		0.0002	ENA-78		0.70
Ig M	+	0.12	LPA	–	0.14	Fibrinogen		0.01	ENRAGE		0.14
MDC	+	0.58	Leptin	+	<0.0001	Haptoglobin		0.05	Endothelin-1		0.52
MIP-1β	+	0.21	SHBG	–	<0.0001	ICAM-1		0.005	Eotaxin	–	0.91
PAPP-A	+	0.04	Tissue Factor	–	0.08	IL-16	–	0.006	Fibrinogen		0.44
TNF RII	+	0.08				Ig E	+	0.08	ICAM-1		0.45
VCAM-1	+	0.03				Insulin	–	0.0005	IL-13		0.47
VEGF	+	0.004				Leptin	–	0.008	IL-12p70		0.73
						MIP-1β		0.07	Ig A	–	0.03
						PAI-1	–	<0.0001	Ig E		0.91
						Stem Cell Factor		0.06	LPA		0.10
						TBG	–	<0.0001	Leptin	+	0.04

Table 2 (continued)

CHD mortality			BMI			HDL-C			Total-C		
Protein	sign	univariate p-value	protein	sign	univariate p-value	Protein	sign	univariate p-value	protein	sign	univariate p-value
						TIMP-1		0.16	MCP-1		<0.0001
						TNF RII		0.007	PAPP-A		0.09
									SGOT	+	0.15
									TNF RII		0.95
									VCAM-1		0.59

Plus and minus signs indicate that, within the multivariate model, the protein is positively respectively inversely related with the endpoint.

Association of proteins with intermediate endpoints

The predictive performance of the newly obtained PLS models for BMI (R^2=0.45), HDL-C (R^2=0.35) and total-C (R^2=0.75) (see figures 2B-D, respectively, and table 2) also improved, as more variance is explained. Leptin was clearly positively related to BMI as well as complement 3 and ferritin, and to a lesser extent ApoB and calcitonin. Furthermore, growth hormone and SHBG were strongly negatively related to BMI, whereas smaller negative effects in the model where found for α-1 antitrypsin, adiponectin, LPA and tissue factor. As HDL-C was negatively correlated with CHD and BMI, HDL-C shows a pattern that is opposite to that of CHD, and more clearly to that of BMI. ApoA1 was strongly positively related to HDL-C. Adiponectin, ApoB, complement 3 and leptin showed an opposite relation with HDL-C compared to BMI. Other proteins related to HDL-C were Ig E (positive) and IL-16, insulin, PAI-1 and TBG (negative). For total-C a strong positive relation was observed with ApoB. Besides ApoB, ApoA1, eotaxin, Ig A, leptin and SGOT were found in relation to total-C levels.

PCA plots: relationships between identified proteins, intermediate endpoints and CHD mortality

PCA analysis was applied to analyze all proteins associated with the different endpoints together with the intermediate endpoints (see figure 3A and 3B for the loading and score plot, respectively). The amount of variance explained by the first three components was equal to 19.2%, 10.8% and 6.6%, respectively. Most proteins that showed to be positively related with CHD mortality in the PLS model are located at the right side of the loading plot (see figure 3A): i.e. Ig M, MDC, VCAM-1, β2-microglobulin, CD40, TNF RII, fibrinogen, CRP, VEGF, MIP-1β and IL-1ra. This corresponds with the score plot (see figure 3B), where more cases are found on the right side of the graph and more controls at the left side. ApoA1 and growth hormone, together with HDL-C, were negatively related to CHD, which also corresponds to the fact that controls are more located at the left side. Although PAPP-A and α-2 macroglobulin were also positively related to CHD in the PLS-model their contribution can however not be clearly interpreted from the PCA plots.

Insulin and leptin, as well as C3 and PAI-1, group together, indicating the positive relationship between these proteins. Corresponding to the PLS results, these proteins show to be positively or negatively related to BMI and HDL-C, respectively. The PCA analysis also confirmed the positive relation of calcitonin and ferritin with BMI, whereas IL-16 and TBG were negatively related to HDL-C, but these relationships were less strong. On the opposite side of the graph SHBG, adiponectin, growth hormone and ApoA1 group together and were strongly negatively or positively related to BMI and HDL-C, respectively. Tissue factor and Ig E, located closely together, were also negatively or positively related to BMI and HDL-C, but to a lesser extent. As the PCA plot shows, ApoB was closely related to total-C, whereas a positive relation with this intermediate endpoint was also the case for SGOT. The positive relationship of leptin with total-C found in the PLS-model was not clearly present in the PCA-plot. Ig A and to a lesser extent eotaxin are located opposite of total-C, which corresponds with the negative relationships found for these proteins in the PLS model of total-C.

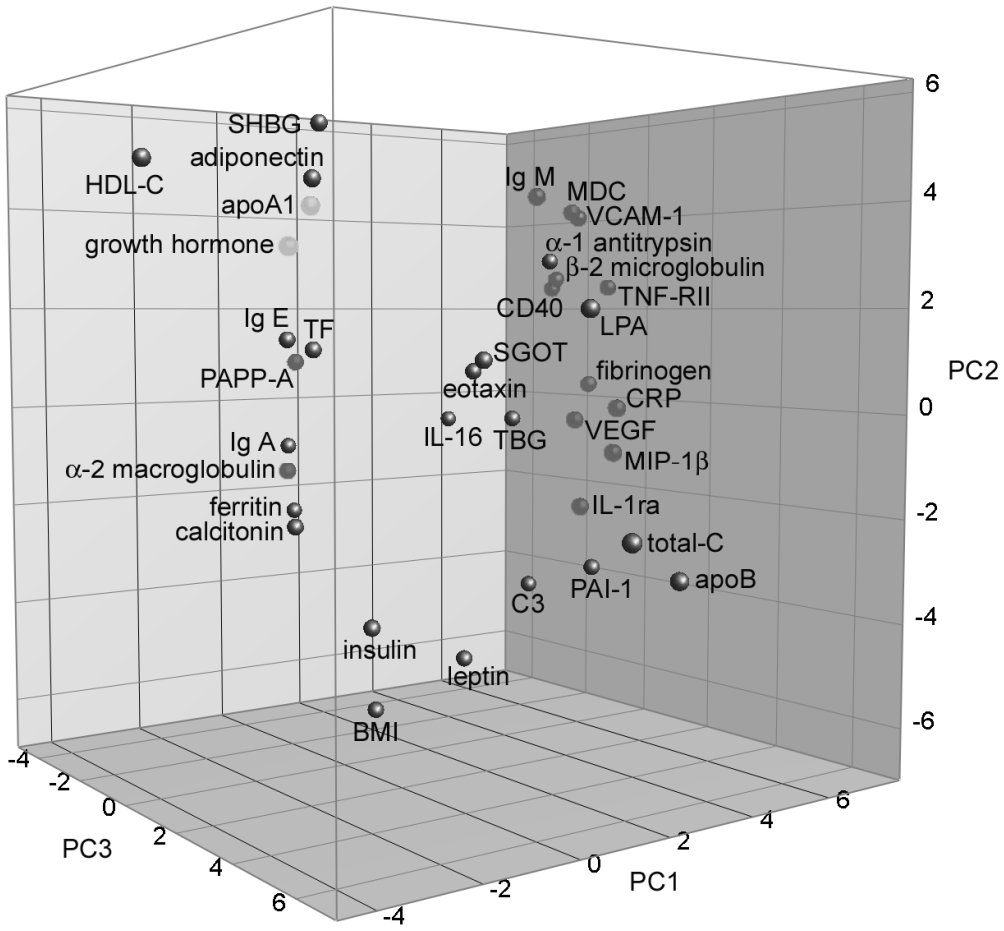


Figure 3A: PCA loading plot including selected proteins and the intermediate endpoints BMI, HDL-C and total-C. In the PCA loading plot the proteins are plotted based on their loadings on the first three principal components. The further proteins are located from the origin, the stronger their contribution in explaining the variance. As most variance is explained by the first component, proteins that have high loadings on the first component contribute most. Proteins and intermediate endpoints grouped together show positive relationships with each other, whereas proteins and intermediate endpoints that are located opposite to each other (as seen from the origin) show negative relationships with each other. Proteins in red and green were respectively positively and negatively associated with CHD mortality. **For full color figure, see page 180.**

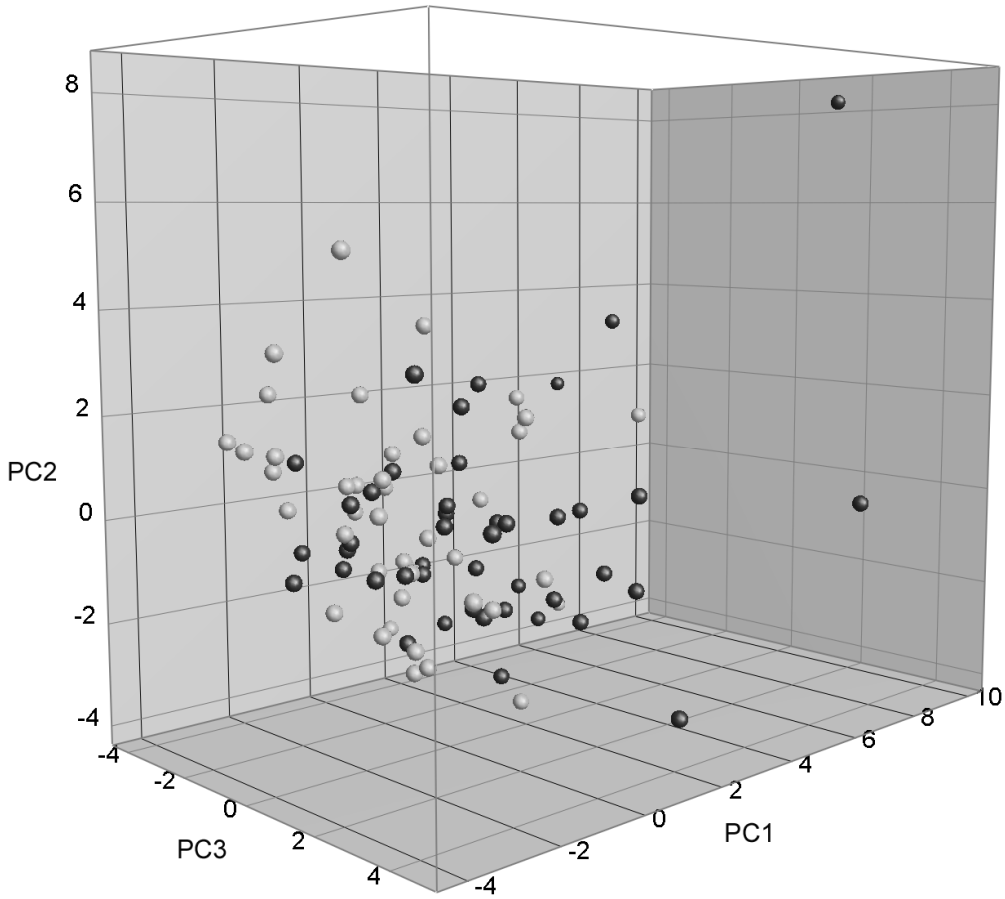


Figure 3B: PCA score plot corresponding to the loading plot. In the PCA score plot the individuals are plotted based on their scores on the first three principal components. Red dots indicate cases of CHD mortality, green dots indicate controls. Observations that are located in the direction of certain proteins score high on these proteins, whereas observations that are located in the opposite direction of certain proteins score low on these proteins (see also figure 3A). **For full color figure, see page 181.**

Discussion

In this study we investigated the association of 83 proteins with CHD mortality, BMI, HDL-C and total-C in men. Besides studying the individual relationships of the proteins with the different endpoints, we applied PLS to conjointly analyze the proteins in relation to the different endpoints, thereby taking relationships between the proteins into account. For CHD mortality a set of 15 proteins was included in the model, which predicted 65 percent of CHD mortality later in life. This set of proteins can serve as a lead for subsequent larger epidemiological studies. Besides validating the results presented in this study, monitoring

the concentrations of this set of proteins in individuals over time could provide useful information about the levels that in an early stage indicate an increased risk for CHD mortality. The proteins associated with CHD mortality in this study do not belong to one functional cluster, but are involved in different biological processes (see table 3). All these processes play a role in atherosclerosis [1, 3, 23, 24].

Applying PLS to select proteins in combination with PCA to visualize the relationships among the proteins and endpoints provided insight in the relationships of proteins and intermediate endpoints with CHD mortality. Groups of proteins involved in inflammation were found to explain most of the variance, as these proteins had high loadings on the first principal component (see PCA loading plot, figure 3A). This result corresponds to the notion of atherosclerosis, which is the main cause of CHD, as an inflammatory disease [24, 25]. Subdividing these inflammatory proteins into functionally related clusters appeared to be difficult; proteins with similar roles were not always projected close to each other. Proteins with high loadings on the second principal component showed to be involved in metabolism. Besides the well-established positive and negative relationships of, respectively, leptin and adiponectin with BMI [20, 26-28], growth hormone and SHBG were also negatively related to BMI and projected closely to adiponectin. Previous studies have shown that growth hormone is inversely related to obesity [29, 30], which may be due to the involvement of this protein in lipolysis [31]. Also, the influence of high insulin levels due to the insulin resistant state that accompanies increased BMI may contribute to the suppression of growth hormone [32].

Table 3: Proteins selected in the PLS-model for CHD and the corresponding biological process in which these proteins are involved.

Protein	Biological process	References
α-2 Macroglobulin	protease activity immune response inflammation lipid metabolism	[33-35]
ApoA1	cholesterol metabolism	[15, 22]
β-2 Microglobulin	immune response inflammation	[36-38]
CD40	immune response inflammation	[39, 40]
CRP	inflammation	[16]
Fibrinogen	coagulation	[41]
Growth hormone	lipid metabolism	[30, 31]
IL-1ra	inflammation	[42, 43]
Ig M	immune response	[44]
MDC	inflammation	[45]
MIP-1β	inflammation	[46, 47]
PAPP-A	inflammation	[48-50]
TNF RII	immune response	[51, 52]
VCAM-1	endothelium/adhesion	[17]
VEGF	angiogenesis	[19]

Furthermore, growth hormone deficient hypopituitary patients were previously shown to have significantly increased BMI, and decreased levels of HDL-C and ApoA1 [53], which supports the results of the PCA loading plot. These patients showed to have a greater absolute risk of a fatal or non-fatal coronary event during the next 5 years compared to controls [53]. Also in our study growth hormone was not only inversely related to BMI, but also to CHD mortality. The inverse association between SHBG levels and BMI has also been found in previous studies [54]. This association may reflect insulin resistance that accompanies increased BMI; insulin was shown to regulate SHBG production with higher insulin levels resulting in reduced SHBG levels [55]. Additionally, complement 3, projected closely to leptin, was also positively related to BMI which is consistent with previous findings from literature [56, 57]. This relationship may be partly mediated by insulin resistance [57]. As BMI is inversely related with HDL-C, adiponectin, ApoB, complement 3 and leptin were also shown to be related to HDL-C. These relationships have also been reported in the literature before [26, 28, 58-60].

Total-C had a high loading on the third component. Proteins with high loadings on the third component include ApoB and Ig A, which are positively and inversely related to total-C, respectively. In the PLS-model for total-C, besides ApoB and Ig A, the proteins eotaxin and SGOT also have a distinct effect. Whereas ApoB is well-known to be related to total-C [21, 61], the relationships of eotaxin, Ig A, and SGOT with total-C are to our knowledge not previously observed in literature and may be a lead for further studies.

Applying the multivariate method PLS provided a more complete view of how groups of proteins relate to the endpoint of interest compared to univariate analysis. The strength of PLS is that, taking other proteins into account, it can detect the importance of proteins which univariately may not seem significant. For example, the individual relationship between growth hormone and CHD mortality was non-significant ($p=0.22$), but the PLS model for CHD mortality clearly showed growth hormone to be of importance. On the other hand, some proteins found to be significant univariately were not selected in the PLS-models. Significant proteins from the univariate analysis have a higher chance to be false positive results; by applying a double cross-validation in the PLS analysis the chance of selecting false positive results is reduced. Another reason may be that the influence of some proteins in the PLS models may be diminished due to the presence of other proteins that have a stronger relationship with the endpoint studied. For example, insulin and BMI were correlated ($r= .44$, $p<0.0001$) in the univariate analysis, but insulin was not associated with BMI in the reduced multivariate PLS-model. Leptin, which is more strongly related to BMI, correlated with insulin ($r= 0.32$, $p<0.002$). Excluding leptin from the PLS analyses resulted in a model very similar to the model with leptin included, but now insulin, and also PAI-1 (in the univariate analysis significantly correlated with BMI, $r=0.21$, $p=0.04$, and leptin, $r=0.32$, $p=0.003$) were additionally included in the model.

PLS is not only able to handle large numbers of variables in moderate to small sample sizes, but also the presence of multicollinearity among the proteins [62]. On the other hand, the application of traditional statistical methods (e.g. logistic regression, multiple linear regression) likely yields unreliable parameter estimates if the number of variables is relatively large compared to the number of observations [63] or when multicollinearity is present [62]. Therefore, to analyze high-dimensional proteomic data PLS is a more suitable tool compared to traditional statistical methods [4].

The threshold to be used for reducing the noise is to some extent subjective. In our PLS-analyses we applied a threshold of RSD<0.5 (i.e. 95% confidence interval of significant effects), but other thresholds could have been chosen. However, in our view, this cut-off was a proper balance between reducing noise while retrieving a sufficient amount of information, providing interpretable PLS-models. Furthermore, proteins that come up as important will not change dramatically for different thresholds.

Conclusions

In this study we applied the multivariate statistical tool PLS to analyze the association between 83 proteins and CHD mortality, BMI, HDL-C and total-C. In this way we identified a set of 15 proteins with prognostic value of CHD mortality later in life. Additionally, sets of proteins were identified to be associated with BMI, HDL-C and total-C. Visualizing the identified proteins together with intermediate endpoints by PCA indicated that proteins involved in inflammation explained most of the variance, followed by proteins involved in metabolism and proteins associated with total-C. Together these results provide a set of proteins with prognostic value for CHD mortality and insight in the relationships among proteins and intermediate endpoints involved in CHD mortality.

Acknowledgement

This project has been carried out in the framework of the Centre for Human Nutrigenomics. The Cardiovascular Disease Risk Factor Monitoring Study (PEILSTATIONS Project) was financially supported by the Ministry of Health, Welfare and Sport of The Netherlands and the National Institute for Public Health and the Environment (RIVM).

References

1. Jain KS, Kathiravan MK, Somani RS, Shishoo CJ: The biology and chemistry of hyperlipidemia. Bioorg Med Chem 2007, 15(14):4674-4699.
2. Bogers RP, Bemelmans WJ, Hoogenveen RT, Boshuizen HC, Woodward M, Knekt P, van Dam RM, Hu FB, Visscher TL, Menotti A, Thorpe RJ Jr, Jamrozik K, Calling S, Strand BH, Shipley MJ: Association of overweight with increased risk of coronary heart disease partly independent of blood pressure and cholesterol levels: a meta-analysis of 21 cohort studies including more than 300 000 persons. Arch Intern Med 2007, 167(16):1720-1728.
3. Link JJ, Rohatgi A, de Lemos JA: HDL cholesterol: physiology, pathophysiology, and management. Curr Probl Cardiol 2007, 32(5):268-314.
4. Boulesteix AL, Strimmer K: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 2007, 8(1):32-44.
5. Jackson JE: A user's guide to principal components. First ed.; New York: Wiley – Interscience; 1991.
6. Verschuren WMM VLE, Blokstra A, Seidell JC, Smit HA, Bueno de Mesquita HB, Obermann-de Boer GL, Kromhout D: Cardiovascular disease risk factors in The Netherlands. Neth J Cardiol 1993, 6:205-210.

7. Kattermann R, Jaworek D, Möller G, Assmann G, Björkhem I, Svensson L, Borner K, Boerma G, Leijnse B, Desager JP, Harwent C, Kupke I, Trinder P: Multicentre study of a new enzymatic method of cholesterol determination. J Clin Chem Clin Biochem 1984, 22:245-251.
8. Lopes-Virella MF, Stone P, Ellis S, Colwell J: Cholesterol determination in high-density lipoproteins separated by three different methods. Clin Chem 1977, 23:882-884.
9. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ: Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 2006, 7:142.
10. Smit S, van Breemen MJ, Hoefsloot HC, Smilde AK, Aerts JM, de Koster CG: Assessing the statistical validity of proteomics based biomarkers. Anal Chim Acta 2007, 592(2):210-217.
11. Stone M: Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological) 1974, 36(2):111-147.
12. Golland P, Liang P, Mukherjee S, Panchenko D: Permutation tests of classification. Lecture Notes in Computer Science 2005, 3559:501-515.
13. Mielke Jr PW, Berry H: Permutation methods: A distance function approach. New York: Springer; 2001.
14. Faber NM: Uncertainty estimation for multivariate regression coefficients Chemometrics and Intelligent Laboratory Systems 2002, 64(2):169-179.
15. Thompson A, Danesh J: Associations between apolipoprotein B, apolipoprotein AI, the apolipoprotein B/AI ratio and coronary heart disease: a literature-based meta-analysis of prospective studies. J Intern Med 2006, 259(5):481-492.
16. de Ferranti SD, Rifai N: C-reactive protein: a nontraditional serum marker of cardiovascular risk. Cardiovasc Pathol 2007, 16(1):14-21.
17. Galkina E, Ley K: Vascular adhesion molecules in atherosclerosis. Arterioscler Thromb Vasc Biol 2007, 27(11):2292-2301.
18. Eaton CB, Gramling R, Parker DR, Roberts MB, Lu B, Ridker PM: Prospective association of vascular endothelial growth factor-A (VEGF-A) with coronary heart disease mortality in Southeastern New England. Atherosclerosis 2008.
19. Khurana R, Simons M, Martin JF, Zachary IC: Role of angiogenesis in cardiovascular disease: a critical appraisal. Circulation 2005, 112(12):1813-1824.
20. Considine RV, Sinha MK, Heiman ML, Kriauciunas A, Stephens TW, Nyce MR, Ohannesian JP, Marco CC, McKee LJ, Bauer TL et al.: Serum immunoreactive-leptin concentrations in normal-weight and obese humans. N Engl J Med 1996, 334(5):292-295.
21. Jungner I, Walldius G, Holme I, Kolar W, Steiner E: Apolipoprotein B and A-I in relation to serum cholesterol and triglycerides in 43,000 Swedish males and females. Int J Clin Lab Res 1992, 21(3):247-255.
22. Dullens SP, Plat J, Mensink RP: Increasing apoA-I production as a target for CHD risk reduction. Nutr Metab Cardiovasc Dis 2007, 17(8):616-628.
23. Blasi C: The autoimmune origin of atherosclerosis. Atherosclerosis 2008.
24. Packard RR, Libby P: Inflammation in atherosclerosis: from vascular biology to biomarker discovery and risk prediction. Clin Chem 2008, 54(1):24-38.
25. Ross R: Atherosclerosis--an inflammatory disease. N Engl J Med 1999, 340(2):115-126.
26. Nishida M, Funahashi T, Shimomura I: Pathophysiological significance of adiponectin. Med Mol Morphol 2007, 40(2):55-67.
27. Arita Y, Kihara S, Ouchi N, Takahashi M, Maeda K, Miyagawa J, Hotta K, Shimomura I, Nakamura T, Miyaoka K, Kuriyama H, Nishida M, Yamashita S, Okubo K, Matsubara K, Muraguchi M, Ohmoto Y, Funahashi T, Matsuzawa Y: Paradoxical decrease of an adipose-specific protein, adiponectin, in obesity. Biochem Biophys Res Commun 1999, 257(1):79-83.
28. Trujillo ME, Scherer PE: Adiponectin--journey from an adipocyte secretory protein to biomarker of the metabolic syndrome. J Intern Med 2005, 257(2)167-175.

29. Brummer RJ: Effects of growth hormone treatment on visceral adipose tissue. Growth Horm IGF Res 1998, 8 Suppl B:19-23.
30. Attallah H, Friedlander AL, Hoffman AR: Visceral obesity, impaired glucose tolerance, metabolic syndrome, and growth hormone therapy. Growth Horm IGF Res 2006, 16 Suppl A:S62-67.
31. Moller N, Gjedsted J, Gormsen L, Fuglsang J, Djurhuus C: Effects of growth hormone on lipid metabolism in humans. Growth Horm IGF Res 2003, 13 Suppl A:S18-21.
32. Luque RM, Kineman RD: Impact of obesity on the growth hormone axis: evidence for a direct inhibitory effect of hyperinsulinemia on pituitary function. Endocrinology 2006, 147(6):2754-2763.
33. Barrett AJ, Starkey PM: The interaction of alpha 2-macroglobulin with proteinases. Characteristics and specificity of the reaction, and a hypothesis concerning its molecular mechanism. Biochem J 1973, 133(4):709-724.
34. Gonzalez P, Alvarez R, Reguero JR, Batalla A, Alvarez V, Cortina A, Cubero GI, Garcia-Castro M, Coto E: Variation in the lipoprotein receptor-related protein, alpha2-macroglobulin and lipoprotein receptor-associated protein genes in relation to plasma lipid levels and risk of early myocardial infarction. Coron Artery Dis 2002, 13(5):251-254.
35. Mocchegiani E, Costarelli L, Giacconi R, Cipriano C, Muti E, Malavolta M: Zinc-binding proteins (metallothionein and alpha-2 macroglobulin) and immunosenescence. Exp Gerontol 2006, 41(11):1094-1107.
36. Shinkai S, Chaves PH, Fujiwara Y, Watanabe S, Shibata H, Yoshida H, Suzuki T: Beta2-microglobulin for risk stratification of total mortality in the elderly population: comparison with cystatin C and C-reactive protein. Arch Intern Med 2008, 168(2):200-206.
37. Wilson AM, Kimura E, Harada RK, Nair N, Narasimhan B, Meng XY, Zhang F, Beck KR, Olin JW, Fung ET, Cooke JP: Beta2-microglobulin as a biomarker in peripheral arterial disease: proteomic profiling and clinical studies. Circulation 2007, 116(12):1396-1403.
38. Saijo Y, Utsugi M, Yoshioka E, Horikawa N, Sato T, Gong Y, Kishi R: Relationship of beta2-microglobulin to arterial stiffness in Japanese subjects. Hypertens Res 2005, 28(6)505-511.
39. Granger DN, Vowinkel T, Petnehazy T: Modulation of the inflammatory response in cardiovascular disease. Hypertension 2004, 43(5):924-931.
40. Lutgens E, Lievens D, Beckers L, Donners M, Daemen M: CD40 and its ligand in atherosclerosis. Trends Cardiovasc Med 2007, 17(4):118-123.
41. Kannel WB: Overview of hemostatic factors involved in atherosclerotic cardiovascular disease. Lipids 2005, 40(12):1215-1220.
42. Arend WP, Malyak M, Guthridge CJ, Gabay C: Interleukin-1 receptor antagonist: role in biology. Annu Rev Immunol 1998, 16:27-55.
43. Tedgui A, Mallat Z: Anti-inflammatory mechanisms in the vascular wall. Circ Res 2001, 88(9):877-87.
44. Matsuura E, Kobayashi K, Tabuchi M, Lopez LR: Oxidative modification of low-density lipoprotein and immune regulation of atherosclerosis. Prog Lipid Res 2006, 45(6):466-486.
45. Gear AR, Camerini D: Platelet chemokines and chemokine receptors: linking hemostasis, inflammation, and host defense. Microcirculation 2003, 10(3-4):335-350.
46. Frangogiannis NG: Chemokines in the ischemic myocardium: from inflammation to fibrosis. Inflamm Res 2004, 53(11)585-595.
47. Liehn EA, Zernecke A, Postea O, Weber C: Chemokines: inflammatory mediators of atherosclerosis. Arch Physiol Biochem 2006, 112(4-5):229-238.
48. Thorn EM, Khan IA: Pregnancy-associated plasma protein-A: an emerging cardiac biomarker. Int J Cardiol 2007, 117(3):370-372.

49. Heeschen C, Dimmeler S, Hamm CW, Fichtlscherer S, Simoons ML, Zeiher AM: Pregnancy-associated plasma protein-A levels in patients with acute coronary syndromes: comparison with markers of systemic inflammation, platelet activation, and myocardial necrosis. J Am Coll Cardiol 2005, 45(2):229-237.
50. Bayes-Genis A, Conover CA, Overgaard MT, Bailey KR, Christiansen M, Holmes DR Jr, Virmani R, Oxvig C, Schwartz RS: Pregnancy-associated plasma protein A as a marker of acute coronary syndromes. N Engl J Med 2001, 345(14):1022-1029.
51. Shai I, Schulze MB, Manson JE, Rexrode KM, Stampfer MJ, Mantzoros C, Hu FB: A prospective study of soluble tumor necrosis factor-alpha receptor II (sTNF-RII) and risk of coronary heart disease among women with type 2 diabetes. Diabetes Care 2005, 28(6):1376-1382.
52. Porsch-Oezcueruemez M, Kunz D, Kloer HU, Luley C: Evaluation of serum levels of solubilized adhesion molecules and cytokine receptors in coronary heart disease. J Am Coll Cardiol 1999, 34(7):1995-2001.
53. Abdu TA, Neary R, Elhadd TA, Akber M, Clayton RN: Coronary risk in growth hormone deficient hypopituitary adults: increased predicted risk is due largely to lipid profile abnormalities. Clin Endocrinol (Oxf) 2001, 55(2):209-216.
54. Key TJ, Allen NE, Verkasalo ZK, Banks E: Energy balance and cancer: the role of sex hormones. Proc Nutr Soc 2001, 60(1)81-89.
55. Pasquali R, Casimirri F, De Iasio R, Mesini P, Boschi S, Chierici R, Flamia R, Biscotti M, Vicennati, V: Insulin regulates testosterone and sex hormone-binding globulin concentrations in adult normal weight and obese men. J Clin Endocrinol Metab 1995, 80(2):654-658.
56. Hernandez-Mijares A, Jarabo-Bueno MM, Lopez-Ruiz A, Sola-Izquierdo E, Morillas-Arino C, Martinez-Triguero ML: Levels of C3 in patients with severe, morbid and extreme obesity: its relationship to insulin resistance and different cardiovascular risk factors. Int J Obes (Lond) 2007, 31(6):927-932.
57. Muscari A, Massarelli G, Bastagli L, Poggiopollini G, Tomassetti V, Drago G, Martignani C, Pacilli P, Boni P, Puddu P: Relationship of serum C3 to fasting insulin, risk factors and previous ischaemic events in middle-aged men. Eur Heart J 2000, 21(13):1081-1090.
58. Mendez-Sanchez N, Gonzalez V, King-Martinez AC, Sanchez H, Uribe M: Plasma leptin and the cholesterol saturation of bile are correlated in obese women after weight loss. J Nutr 2002, 132(8):2195-2198.
59. Sniderman AD, Faraj M: Apolipoprotein B, apolipoprotein A-I, insulin resistance and the metabolic syndrome. Curr Opin Lipidol 2007, 18(6):633-637.
60. Ylitalo K, Pajukanta P, Meri S, Cantor RM, Mero-Matikainen N, Vakkilainen J, Nuotio I, Taskinen MR: Serum C3 but not plasma acylation-stimulating protein is elevated in Finnish patients with familial combined hyperlipidemia. Arterioscler Thromb Vasc Biol 2001, 21(5):838-843.
61. Leino A, Impivaara O, Kaitsaari M, Jarvisalo J: Serum concentrations of apolipoprotein A-I, apolipoprotein B, and lipoprotein(a) in a population sample. Clin Chem 1995, 41(11):1633-1636.
62. Tobias RD: An introduction to Partial Least Squares Regression. SAS Institute: Cary, NC, 1997.
63. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996, 49(12):1373-1379.

Chapter 8

General discussion

Introduction

Possibilities and limitations in analyzing large nutrigenomic datasets in relation to complex diseases

Benefits and limitations of genomic research for public health

Future perspective

Introduction

Genes and proteins play a role in the development of complex diseases in humans by complex interactions. Therefore it is of importance to take these interactions into account in genomic studies. In this thesis, approaches to analyze genomic datasets to identify these complex relationships have been discussed and applications to real datasets have been shown. In the general discussion the results of the studies performed will be placed within the discussion of the possibilities and limitations currently present in the analysis of large nutrigenomic datasets. This is followed by the discussion of the benefits and limitations of genomic research for public health. In the last section a future perspective on nutrigenomic research will be discussed.

Possibilities and limitations in analyzing large genomic datasets in relation to complex diseases

In this section possibilities and limitations in the analysis of genomic datasets will be discussed for the aim of obtaining insight in the biological mechanisms and for the aim of diagnosis and prognosis.

Insight in the biological mechanisms

For the aim of obtaining insight in the biological mechanisms, possibilities and limitations in the analysis will be discussed according to the points of the multi-step approach that was presented in chapter 1: detection of heterogeneity, dimensionality reduction, statistical interpretation and biological interpretation.

Detection of heterogeneity

In chapter 2 an overview of the strengths and weaknesses of multi-locus methods to handle statistical challenges in the analysis of large numbers of SNPs was provided. Besides other issues, the performance of the multi-locus methods in the presence of heterogeneity was discussed. As indicated in this chapter, methods that test the association between predictors and disease for the total sample will be affected by the presence of heterogeneity, whereas neural networks and recursive partitioning methods are assumed to be able to handle heterogeneity [1, 2]. However, contradictory findings for the ability of recursive partitioning methods to handle heterogeneity were found in different simulation studies [3, 4]. Further work is needed to investigate the conditions whereby recursive partitioning methods are able to detect heterogeneity. Besides neural networks and recursive partitioning methods, other methods can be applied to handle heterogeneity [5, 6], including cluster methods [7]. If different clusters have been obtained by cluster analysis, subsequent analyses should be performed separately for each cluster (stratified analyses) or the cluster variable should be included in the analysis.

Dimensionality reduction

For dimensionality reduction, first the approach of prioritization and selection will be discussed. The methods discussed in chapter 2 were previously shown to have good power

to detect true associations [4, 8-10]. As these methods have different strengths and weaknesses in the prioritization and selection of important SNPs, the main conclusion of this chapter was that the application of a combination of methods is likely a useful strategy to analyze SNPs in relation to complex diseases. This is underscored by a simulation study in which different multi-methods were compared in their performance to select SNPs and detect interactions between SNPs from simulated data [3]. As indicated in chapter 3 of this thesis, application of different multi-locus methods to select SNPs from real data showed that this combined analytical strategy indeed has several advantages compared to applying only one method. It should be noted that more methods have become available that can be applied for analysis of large SNP datasets, but which are not addressed in this thesis, e.g. [11-13].

Some of the methods discussed in chapter 2 were specifically developed for the analysis of SNP data, to identify important SNPs and SNP combinations in relation to traits and complex diseases. Other methods included in this chapter can also be applied to microarray and proteomic data, such as neural networks [14] and random forests (RF) [15, 16]. RF is a suitable method to capture all possible interactions in the prioritization of variables, and was therefore applied in chapter 3, 5 and 6 for prioritization of SNPs, genes and mass/charge ratios, respectively. It is useful to analyze data that consists of either continuous predictors, or categorical predictors consisting of similar number of categories. However, caution must be taken if categorical predictors consisting of different number of categories or both categorical and continuous predictors are analyzed simultaneously. In that case the application of RF will likely lead to biased results and conditional inference trees can be used in the forest [17, 18]. Another issue is that RF does not provide a threshold to denote which variables should be selected for further statistical and biological interpretation. In this thesis we investigated several ways to define the threshold, including robustness of the ranking of variables (chapter 3) and a permutation approach (chapter 5). Another way would be to screen the top-ranked variables for biological relevance. However, this is not preferred, as biological information is not always known beforehand and relevant information unknown to the researcher could be disregarded.

Of course, applying a combination of different methods will also likely be a viable approach to analyze microarray data and proteomic data. Methods to select a subset of genes from microarray data have been compared by others, but these methods did not take interactions into account [19]. Therefore, the next step will be to compare methods that take interactions into account to select important variables in the analyses of microarray and proteomic data.

The approach of reducing the dimensionality by methods that combine variables into a smaller set of new variables will now be discussed. In this respect, PLS is a suitable tool to analyze high-dimensional genomic data [20]. In chapter 7 we applied partial least squares (PLS) to analyze the association of 83 plasma proteins with CHD mortality, body mass index (BMI), HDL-cholesterol (HDL-C), and total cholesterol (total-C). This study points out that the noise of non-important variables may obscure the signals of important variables. Therefore, if statistical results are obtained that are biologically plausible, but not statistically significant, a reduction in the number of variables to reduce the noise is an option. PLS may not always be able to detect variables with small effects [21] and therefore it may not retrieve all biologically relevant information from the data. In chapter 7,

application of PLS indeed showed that proteins with smaller effects were not always selected. As indicated in this chapter, a reason for this may be that these proteins were related to another protein that showed to have a larger effect on the endpoint studied.

Statistical interpretation

As indicated above, applying a combination of methods can help in the selection of SNPs, but such a combined analytical strategy can also help for statistical interpretation to understand how variables contribute to the studied disease. Similarities and differences in results between methods provide information on the type of effects that are found (see for example chapter 3). Also, comparing advanced methods that take interaction into consideration with univariate results provides insight whether variables selected in a multivariate model contribute by themselves and/or in interaction with other variables. For example, although RF does take all possible interactions into account, it does not provide information on whether predictors contribute by their main effect or in interaction, and therefore it is useful to compare the ranking by RF with the ranking by a univariate test, as shown in chapters 3 and 5.

Visualization tools are available that show graphically the importance of individual variables as well as the interactions between variables. The visualization tools that were applied for statistical interpretation were the interaction entropy graph (see chapters 3 and 4) and the interaction dendrogram (see chapter 3) [22-24]. Besides these visualization tools, the parametric method logistic regression analysis was applied. The application of logistic regression analysis is valid at this step if the number of observations is sufficiently large compared to the number of selected variables [25]. The associations that were found by the interaction entropy graph in both chapters 3 and 4 were confirmed by logistic regression analysis. This indicates that the interaction entropy graph is a useful tool for statistical interpretation of a selected subset of variables. Another visualization tool (not discussed in this thesis) that can be applied for statistical interpretation of SNP data is Bayesian networks [26].

The interaction dendrogram that was applied in chapter 3 is a hierarchical clustering method. Besides this method, other cluster methods can be applied to find among the selected variables groups of variables that statistically are closely related. In chapter 5 we used for statistical interpretation the cluster method self-organizing maps (SOM) [27] to find groups of clusters with similar gene expression patterns. Clusters containing genes exclusively selected by RF but not by t-test were assumed to be involved by their interaction. These clusters were shown to have biological meaning. SOM was chosen above other cluster methods as it has the advantage that it provides an ordering in the clusters, which is very useful in the context of large numbers of clusters. Still, different cluster algorithms have different strengths and weaknesses in clustering genes by their expression profile into functional groups; therefore application of several cluster methods and evaluation by existing biological information has been recently recommended [28].

In chapter 7, after the selection of proteins by PLS, we applied for further statistical interpretation principal component analysis (PCA). Combining PLS with PCA seems a useful approach to obtain an overview of the proteomic data; in our study this approach provided insight in the relationships among identified proteins, intermediate endpoints and CHD mortality.

Biological interpretation

To improve biological interpretation for a set of selected important genes or proteins, pathway analysis can be helpful to provide a more global overview of what biological processes are significantly related to the endpoint of interest. Different pathway analysis tools are available, including methods that import a subset of preselected genes and assess enrichment of biological processes [29, 30] and gene set analysis methods that circumvent the need to preselect genes [31]. However, pathway analysis programs currently have the limitations to take only those genes into account that are annotated in the database used by the program. Therefore, in order to overcome these limitations, we developed in chapter 5 a framework to analyze microarray data in which both gene ontology (GO)-annotated and non-annotated genes are taken into account. In this chapter it is shown that more than half of the microarray data is not included if pathway analysis programs are applied that are solely based on functional annotation and co-occurrence in gene sets. Applying this framework taking both GO-annotated and non-annotated genes into account, more biological information could be retrieved from the data compared to gene set enrichment analysis (GSEA), indicating that it is important to include all microarray data in the analysis. For biological interpretation, data mining and literature search were also performed to obtain more insight.

Within genomics, besides the study of the different biological levels by genetics, transcriptomics and proteomics, a systems biology approach is nowadays more and more emerging [32]. For systems biology, which is also referred to as integrative biology [33], many different definitions are available [34]. One way to define systems biology is as a holistic approach for studying biological systems that analyzes multiple macromolecular species (e.g. DNA polymorphisms, RNA, protein, metabolites) in one experiment [35]. In the systems biology approach, data of the different omics techniques is integrated and a combined analytical approach is applied for interpretation [36]. Now genetics, transcriptomics and proteomics are more and more established, integrating the information from the different biological levels is a logical and important next step. Integration of information from different omics techniques will help to reduce false positive and false negative results obtained from single omics approaches. Furthermore, this approach enables researchers to combine results from different biological levels to improve biological interpretation and construct a more all-inclusive biological model. Systems biology has shown to have advantages in finding underlying biological patterns [36].

Summary of the multi-step approach

For the different points considered within the multi-step approach to obtain insight in biological mechanisms, a summary of the results obtained within the studies performed in this thesis are included in the grey boxes in figure 1.

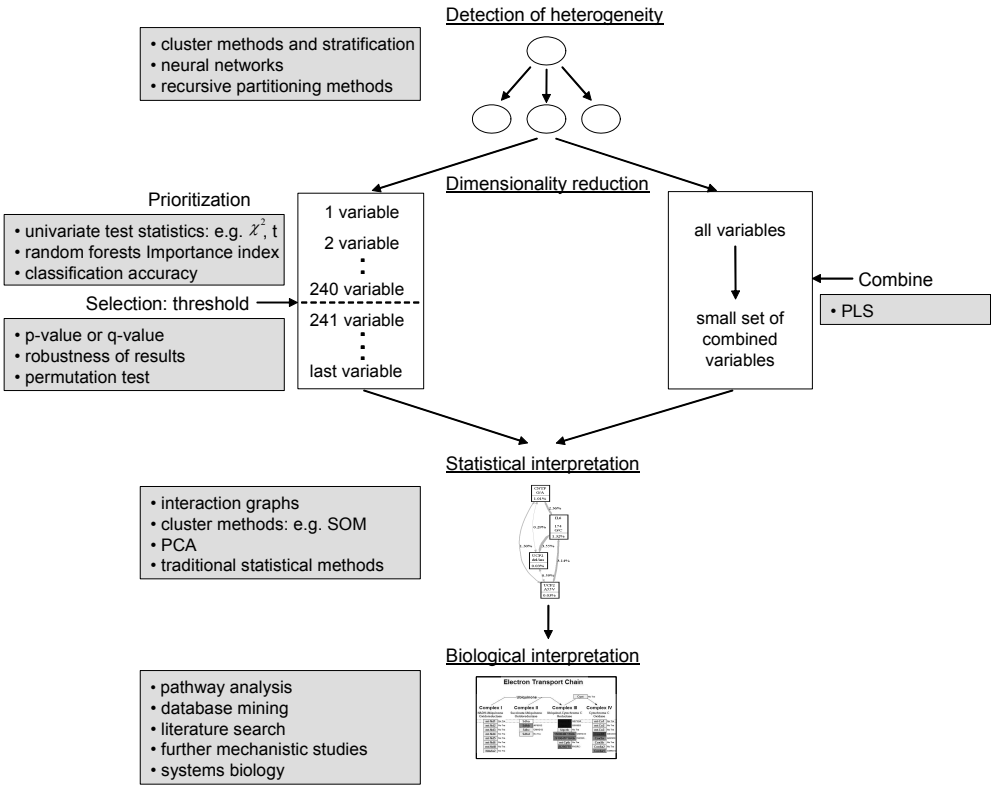


Figure 1: Multi-step approach to analyze large genomic datasets. For each step, grey boxes include methods that can be used for this specific step.

Diagnosis and prognosis

Based on the results obtained from the studies performed in this thesis, first the possibilities and limitations in the analysis for the aim of diagnosis will be discussed, followed by the discussion of the analysis for the aim of prognosis.

Diagnosis

In chapter 6 a three-step approach to develop a discrimination rule for classification of breast cancer patients and control subjects was described. This endeavour was part of a classification competition in which 10 research groups applied each their preferred approach. The objective of the competition was to see what approach would yield the best classification and prediction accuracy to discriminate between breast cancer patients and control subjects. This classification competition has been described in a special journal issue [37]. In chapter 6 we developed our approach on the training dataset, which yielded a good classifying model with a sensitivity and specificity of 86.8% and 85.7%, respectively.

Approaches that have been applied by other groups include RF, RF in combination with linear discriminant analysis, principal components discriminant analysis, support vector machine, logistic ridge regression (see table 1). The advantage of the three-step approach we applied is that our model can be simply interpreted, which is hard to do with for example a forest model or a model obtained by SVM. Valkenborg et al. [38], one of the participating research groups, first compared different classification techniques and found that RF and SVM yielded the best classification. The classification of SVM was only marginally better, therefore the classification results based on SVM were reported (see table 1). For microarray data it has also been shown previously that RF, SVM, but also diagonal linear discriminant analysis yielded good classification results [39]. The disadvantage of SVM is that it is not straightforward to extend this method to more than two classes, whereas RF can be easily applied to multi-class problems. For RF a variable selection approach for classification purposes has been developed that can be applied to select genes from microarray data [15, 40].

As can be seen from table 1, one of the outcomes of the classification competition was that different approaches yielded different rules with similar performance. The finding that different subsets of variables may have equal performance in classifying cases and controls is more often observed in applying different classification techniques [41, 42]. However, this may also depend on the type of data; mass spectrometry data often contains more variables with large effects, yielding different possible solutions with equal classification performance.

On the other hand, classification of a complex disease using SNP data appears to be more problematic. In a separate study we also investigated whether we could improve the prediction of a binary outcome using SNP data by a modified approach of MDR. This modified approach was based on the idea that there is likely more than one combination of SNPs involved in complex diseases and that by predicting the outcome variable based on several SNP combinations we could improve the prediction compared to the prediction based on only one SNP combination. In this modified approach, the total dataset was split in a training set and a test set. MDR was applied to analyze the training set in order to select the most important single SNPs and 2-SNP combinations. Subsequently, for each of these top single SNPs and 2-SNP combinations dummy variables were created based on the risk group assigned to each genotype or multi-locus genotype by MDR: low risk genotypes were coded as 0 and high risk genotypes as 1. These dummy variables were included in a stepwise manner in the logistic regression model to predict the case status for the individuals in the test set. In this way a prediction accuracy based on several single SNPs and 2-SNP combinations was obtained. However, only a marginal improvement in the prediction accuracy was observed compared to the best SNP combination (results not shown) and therefore this approach was not further developed. Thus, obtaining good predictions of complex diseases based on SNP data is often difficult to achieve. This is one of the current limitations of genomic research for public health, as is also discussed below.

Table 1: Comparison of the different approaches that were applied to correctly classify cases of breast cancer and controls. The approaches are ranked by their total accuracy on the training dataset.

Approach	Sensitivity (%)	Specificity (%)	Classification accuracy (%)	Reference
RF	81.6	85.7	83.7	[43]
PCDA	81.6	85.7	83.7	[44]
SVM	81.6	87.0	84.3	[38]
PCDA	82.9	89.6	86.3	[45]
Three-step approach	86.8	85.7	86.3	Chapter 6
RF+LDA	90.8	84.4	87.6	[46]
Empirical bayes LR	88.2	89.6	88.9	[47]
Autocorrelated LRR	89.5	89.6	89.5	[48]
Conformal predictor (based on SVM)	89.5	92.2	90.8	[49]
SVM	89.5	98.7	94.1	[50]

PCDA: principal component discriminant analysis
LR: Logistic regression
LRR: Logistic ridge regression
SVM: Support vector machine

Prognosis

In chapter 7 we applied PLS to identify a set of proteins with prognostic value for CHD mortality later in life. As PLS iteratively maximizes the covariance between the latent components and the response variable, PLS is advantageous in prediction problems. CHD cases died on average six years after baseline, and a PLS model including a set of 15 proteins was identified that predicted 65% of CHD mortality later in life. We also investigated whether we could predict the number of years cases died after baseline examination based on the total set of proteins. However, it was not possible to obtain a PLS model with good predictive performance.

The selection of the 15 proteins was based on a cut-off of a relative standard deviation (RSD) ≤ 0.5, which is a valid approach to select variables. We did not formalize the application of PLS as a variable selection method to obtain the smallest set of proteins that would yield the best prediction. Other possible ways for variable selection can be used, among others, a recursive feature elimination (RFE) procedure (similar to the parameter decreasing method (PDM) discussed in chapter 2), uninformative variable elimination and a genetic algorithm [51]. Variable selection strategies for classification are subject for further investigation.

Principal components regression (PCR) [52, 53] is a technique related to PLS. The first step in PCR is to perform PCA to construct the components in order to maximize the variance explained in a set of predictor variables. This is followed by a multiple linear regression step whereby the scores obtained by PCA are related to the outcome. Thus, in PCR the components are constructed without taking the covariance structure with the outcome variable into account. Therefore PCR may yield less predictive models than PLS.

Benefits and limitations of genomic research for public health

Human genetic/genomics research has shown to have an impact for Mendelian disorders [54], but not yet for complex diseases. The number of genes that have been really identified so far in relation to complex diseases is still limited and only very few evidence-based guidelines in genetics have been made [55]. BRCA mutation testing for breast cancer susceptibility is one of the genetic/genomic applications that could be used according to recommendations of U.S. Preventive Services Task Force (USPSTF) [56]. As discussed above, the predictive value of SNPs is limited for complex diseases. Also in our study performed in chapter 3 the prediction accuracy of SNPs in relation to dichotomized HDL was limited for the models obtained by RF (55%) and MDR (57.9%). This is due to the many genetic, environmental and social factors involved in intermediate endpoints and complex diseases; most SNPs will therefore have a low predictive value and associated attributable risk and the clinical application of SNPs is questionable [57]. For example, the predictive performance of SNPs is lower compared to other risk factors [58]. To identify individuals with an increased risk of a complex disease, family history might be more useful as a screening tool [59]. Subsequently, personalized prevention strategies could be provided to these individuals.

Another limitation of most genetic association studies that have been performed so far is that replication of associations between SNPs and complex diseases in independent studies has often failed [60, 61]. This is due to several reasons, including technical issues regarding genotyping [62], heterogeneous populations [6, 63], the modest to small sample sizes resulting in lack of power [64] and false positive results, and absence of performing multivariate statistical analyses to investigate complex interactions. Conducting meta-analyses can be helpful in determining the robustness of the results of reported associations between genes and disease [65], see for example [66]. However, although useful, these meta-analyses are also subject to the issues present in genetic association studies [67]. For example, combining data from different studies may increase heterogeneity and thereby dilute true associations. Also, meta-analyses may miss true associations that are present due to gene-gene and gene-environment interactions. Therefore meta-analyses should be supplemented with large epidemiological studies that apply approaches such as discussed in this thesis to investigate these interactions. The use of this type of complex information for health care is only starting to be discussed [55].

Despite the current limitations of the contribution of genomic research for public health, the application of genomics can provide the opportunity to improve public health. For example, identifying biomarkers for diagnostic and prognostic tests may be used in the clinical setting. In this respect, the results of chapter 6 and 7 show the value of genomic research for diagnosis and prognosis, respectively, of complex diseases. Furthermore, with the current development in technology and possibilities to measure large numbers of variables for a decreasing price, epidemiological studies including many factors are more and more feasible. The study performed in chapter 7 is unique as it is one of the first studies in which the concentrations of large numbers of plasma proteins are investigated in relation to CHD and intermediate endpoints. In contrast to univariate statistics, applying multivariate statistics to take the simultaneous effects of many proteins into account yielded interpretable models for CHD, BMI, HDL- and total cholesterol, that provided more

complete information. This information is therefore more useful for development of possible treatment procedures. Also, genome wide association studies are a first step to genomic applications for complex diseases in humans and evidence from the first large genome wide association studies are promising [68-71]. As discussed in this thesis, approaches (e.g. RF) are nowadays available that are useful to analyze the large amount of data obtained in these studies and have the advantage of taking interactions into account (see also [72]). These new advances might bring discoveries that are useful for public health.

Effective translation of genomic research into improved public health is needed. Recently, a framework was described consisting of four phases for integration of results from genomic studies into health care and disease prevention, viz. 1) from gene discovery to candidate health applications; 2) from health application to evidence-based guidelines; 3) from evidence-based guidelines to health practice; and 4) from practice to population health impact [55]. Most of the human genetics and genomics articles (>350,000) have been published on the first phase, while less than 3% deals with translation research from phase 2 and beyond [55]. If results from genomic research applications in public health yield applications and evidence-based guidelines it is also important to evaluate how well this information will translate into better prevention, diagnosis, and treatment. At the moment there is a gap in translation of evidence-based guidelines into health practice that is especially present in preventive medicine [55, 73], which is the focus of nutrigenomic research. To evaluate the contribution of genomic applications to public health, research should integrate studies of health outcomes with studies of ethical, legal and social implications [57]. Genomic research has raised many ethical issues on how to deal with information obtained from genomic research, e.g. how to deal with incidental findings in genomic research [74], genetic information of individuals that reveals information about other family members [75], implications of genetic testing for discrimination of health insurance [75, 76] and the potential harm (e.g. increased anxiety, unnecessary treatment) and benefits of genetic testing [77-79]. In contrast to other applications of genomic research where the focus is on discerning patients from non-diseased individuals, nutrigenomic research is directed to public health and prevention of disease shifting the focus from patients to apparently healthy individuals, which raises specific ethical issues [80, 81]. At the moment ethics is still separate from genomic research and it is important that ethics will become an integral part of genomic research [57].

Future perspectives

In this thesis approaches have been discussed to analyze genetic polymorphisms, mRNA levels, peptide and protein concentrations in relation to complex diseases in humans. As discussed, the next step will be to combine genetic, transcriptomic and proteomic data in a systems biology approach. This may in the near future show to be useful to generate more global and comprehensive biological hypotheses regarding the development of complex diseases in humans, but also for the detection of biomarkers for clinical purposes. Also, validation of effects on different biological levels will reduce the number of false positive and false negative results. For statistical analyses of genomic datasets an important aspect that will increase the progress within genomic research is the free availability of easy-to-use

software to perform the analyses. The analyses that have been performed in this thesis have been mainly performed with freely available software, and useful sources for free software include R-packages [82], MDR [83], Orange software [84], and Weka [85]. Computer memory and calculation performance at the moment are also a bottleneck to easily perform analyses, limiting the possibility for exhaustive searches over all possible combinations of variables. In our studies we also made use of a computer cluster to be able to perform the RF analyses with large numbers of trees, which was not feasible on a normal PC. Increase in computer memory and calculation performance will also move the field of genomics forward, providing the possibility to detect intricate interactions among large numbers of variables.

In genomics, studying SNPs, mRNA levels and proteins in relation to complex diseases is nowadays a joined effort of researchers from different disciplines, e.g. epidemiologists, statisticians, bioinformaticians and biologists. For genomics research to become successful, besides integration of information from different omic techniques, communication and collaboration between different research groups is important. Large genomic research networks exist [86] that facilitate cooperation between different research groups. Multidisciplinary and interdisciplinary [57] nutrigenomic research would also be enhanced by including the disciplines of public health, ethics, behavioral and social sciences [55]. Hopefully the large nutrigenomic research networks can extend to form a global scientific community to make an effort to reduce the prevalence of disease, develop prevention strategies and enhance the health and well-being of human beings.

Finally, besides providing evidence leading to improvements of health and well-being, it is the responsibility of biomedical scientists to increase public awareness on health and health-related technologies, especially in the low socio-economic group. With respect to genomics, promoting genetic literacy of both the public and health professionals [57] and communicating the complexity of genetic risk are of importance [77, 87, 88]. Besides communication of genetic information, policy should also continue to be directed at changes in environmental factors that contribute to public health. Policy directed at smoking cessation has been successful over the last decades, but obesity reverses this gain in public health [89]. Therefore, similar advances as made in smoking cessation are needed in promoting recommendations for healthy life-style behavior including healthful nutrition [90], to not only reduce risk factors and the prevalence of corresponding complex diseases, but also to improve human well-being.

References

1. Province MA, Shannon WD, Rao DC: Classification methods for confronting heterogeneity. Adv Genet 2001, 42:273-286.
2. Lucek PR, Ott J: Neural network analysis of complex traits. Genet Epidemiol 1997, 14(6):1101-1106.
3. Vermeulen SH, Den Heijer M, Sham P, Knight J: Application of multi-locus analytical methods to identify interacting loci in case-control studies. Ann Hum Genet 2007, 71(Pt 5):689-700.
4. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004, 5(1):32.

5. Thornton-Wells TA, Moore JH, Haines JL: Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet 2004, 20(12):640-647.
6. Hu D, Ziv E: Confounding in genetic association studies and its solutions. Methods Mol Biol 2008, 448:31-39.
7. Thornton-Wells TA, Moore JH, Haines JL: Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data. BMC Bioinformatics 2006, 7:204.
8. Motsinger AA, Lee SL, Mellick G, Ritchie MD: GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics 2006, 7:39.
9. Ritchie MD, Hahn LW, Moore JH: Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003, 24(2):150-157.
10. Wille A, Hoh J, Ott J: Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. Genet Epidemiol 2003, 25(4):350-359.
11. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD: Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol 2008, 32(4):325-340.
12. Gayan J, Gonzalez-Perez A, Bermudo F, Saez ME, Royo JL, Quintas A, Galan JJ, Moron FJ, Ramirez-Lorca R, Real LM et al: A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. BMC Genomics 2008, 9(1):360.
13. Schwender H, Ickstadt K: Identification of SNP interactions using logic regression. Biostatistics 2008, 9(1):187-198.
14. O'Neill MC, Song L: Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC Bioinformatics 2003, 4:13.
15. Diaz-Uriarte R, Alvarez de Andres S: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006, 7:3.
16. Breiman L: Random forests. Machine learn 2001, 45:5-32.
17. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007, 8:25.
18. Hothorn T, Hornik K, Zeileis A: Unbiased recursive partitioning: a conditional inference framework. J Comput Graphical Stat 2006, 15(3):651-674.
19. Jeffery IB, Higgins DG, Culhane AC: Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 2006, 7:359.
20. Boulesteix AL, Strimmer K: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 2007, 8(1):32-44.
21. Tobias RD: An introduction to Partial Least Squares Regression. SAS Institute: Cary, NC, 1997.
22. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 2006, 241(2):252-261.
23. Jakulin A, Bratko I: Analyzing attribute dependencies. Lect Notes Artif Intell 2003, 2838:229-240.
24. Jakulin A, Bratko I, Smrke D, Demsar J, Zupan B: Attribute interactions in medical data analysis. Lect Notes Artif Intell 2003, 2780:229-238.
25. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996, 49(12):1373-1379.
26. Rodin A, Mosley TH, Jr., Clark AG, Sing CF, Boerwinkle E: Mining genetic epidemiology data with Bayesian networks application to APOE gene variation and plasma lipid levels. J Comput Biol 2005, 12(1):1-11.

27. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 1999, 96(6):2907-2912.
28. Do JH, Choi DK: Clustering approaches to identifying gene expression patterns from DNA microarray data. Mol Cells 2008, 25(2):279-288.
29. Khatri P, Draghici S: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005, 21(18):3587-3595.
30. Rivals I, Personnaz L, Taing L, Potier MC: Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2007, 23(4):401-407.
31. Nam D, Kim SY: Gene-set approach for expression pattern analysis. Brief Bioinform 2008, 9(3):189-197.
32. Westerhoff HV, Palsson BO: The evolution of molecular biology into systems biology. Nat Biotechnol 2004, 22(10):1249-1252.
33. Liu ET: Systems biology, integrative biology, predictive biology. Cell 2005, 121(4):505-506.
34. Lisacek F, Appel RD: Systems Biology: a loose definition. Proteomics 2007, 7(6):825-827.
35. Kaput J, Rodriguez RL: Nutritional Genomics: Discovering the Path to Personalized Nutrition. Hoboken: John Wiley &Sons, Inc.; 2006.
36. Ge H, Walhout AJ, Vidal M: Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet 2003, 19(10):551-560.
37. Competition on Clinical Mass Spectrometry Based Proteomic Diagnosis. Statistical Applications in Genetics and Molecular Biology 2008, 7(2).
38. Valkenborg D, Van Sanden S, Lin D, Kasim A, Zhu Q, Haldermans P, Jansen I, Shkedy Z, Burzykowski T: A cross-validation study to select a classification procedure for clinical diagnosis based on proteomic mass spectrometry. Stat Appl Genet Mol Biol 2008, 7:Article12.
39. Van Sanden S, Lin D, Burzykowski T: Performance of classification methods in a microarray setting: a simulation study. Communications in Statistics - Simulation and Computation 2007, 37:418-433.
40. Diaz-Uriarte R: GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 2007, 8:328.
41. Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA: Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 2004, 32(2):71-83.
42. Huang X, Pan W, Grindle S, Han X, Chen Y, Park SJ, Miller LW, Hall J: A comparative study of discriminating human heart failure etiology using gene expression profiles. BMC Bioinformatics 2005, 6:205.
43. Barrett JH, Cairns DA: Application of the random forest classification method to peaks detected from mass spectrometric proteomic profiles of cancer patients and controls. Stat Appl Genet Mol Biol 2008, 7:Article4.
44. Hoefsloot HC, Smit S, Smilde AK: A classification model for the Leiden proteomics competition. Stat Appl Genet Mol Biol 2008, 7:Article8.
45. Fearn T: Principal component discriminant analysis. Stat Appl Genet Mol Biol 2008, 7:Article6.
46. Datta S: Classification of breast cancer versus normal samples from mass spectrometry profiles using linear discriminant analysis of important features selected by random forest. Stat Appl Genet Mol Biol 2008, 7:Article7.
47. Strimenopoulou F, Brown PJ: Empirical Bayes logistic regression. Stat Appl Genet Mol Biol 2008, 7:Article9.
48. Goeman JJ: Autocorrelated logistic ridge regression for prediction based on proteomics spectra. Stat Appl Genet Mol Biol 2008, 7:Article10.

49. Gammerman A, Nouretdinov I, Burford B, Chervonenkis A, Vovk V, Luo Z: Clinical mass spectrometry proteomic diagnosis by conformal predictors. Stat Appl Genet Mol Biol 2008, 7(2):Article13.
50. Pham TV, van de Wiel MA, Jimenez CR: Support vector machine approach to separate control and breast cancer serum samples. Stat Appl Genet Mol Biol 2008, 7:Article11.
51. Czekaj T, Wu W, Walczak B: Classification of genomic data: some aspects of feature selection. Talanta 2008, 76: 564-574.
52. Jollife IT: A note on the use of principal components in regression. Applied Statistics 1982, 31:300-303.
53. Hadi AS, Ling RF: Some Cautionary Notes on the Use of Principal Components Regression. The American Statistician 1998, 52:15-19.
54. Jimenez-Sanchez G, Childs B, Valle D: Human disease genes. Nature 2001, 409(6822):853-855.
55. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L: The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet Med 2007, 9(10):665-674.
56. Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement. Ann Intern Med 2005, 143(5):355-361.
57. Burke W, Khoury MJ, Stewart A, Zimmern RL: The path from genome-based research to population health: development of an international public health genomics network. Genet Med 2006, 8(7):451-458.
58. Gail MH: Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst 2008, 100(14):1037-1041.
59. Yoon PW, Scheuner MT, Khoury MJ: Research priorities for evaluating family history in the prevention of common chronic diseases. Am J Prev Med 2003, 24(2):128-135.
60. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE et al: Replicating genotype-phenotype associations. Nature 2007, 447(7145):655-660.
61. Redden DT, Allison DB: Nonreplication in genetic association studies of obesity and diabetes research. J Nutr 2003, 133(11):3323-3326.
62. Pompanon F, Bonin A, Bellemain E, Taberlet P: Genotyping errors: causes, consequences and solutions. Nat Rev Genet 2005, 6(11):847-859.
63. Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, Suk EK, Toliat MR, Klopp N, Caliebe A et al: SNP-based analysis of genetic substructure in the German population. Hum Hered 2006, 62(1):20-29.
64. Klein RJ: Power analysis for genome-wide association studies. BMC Genet 2007, 8:58.
65. Munafo MR, Flint J: Meta-analysis of genetic association studies. Trends Genet 2004, 20(9):439-444.
66. Thompson A, Di Angelantonio E, Sarwar N, Erqou S, Saleheen D, Dullaart RP, Keavney B, Ye Z, Danesh J: Association of cholesteryl ester transfer protein genotypes with CETP mass and activity, lipid levels, and coronary risk. Jama 2008, 299(23):2777-2788.
67. Salanti G, Sanderson S, Higgins JP: Obstacles and opportunities in meta-analysis of genetic association studies. Genet Med 2005, 7(1):13-20.
68. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007, 447(7145):661-678.
69. Lango H, Weedon MN: What will whole genome searches for susceptibility genes for common complex disease offer to clinical practice? J Intern Med 2008, 263(1):16-27.

70. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE et al: Genomewide association analysis of coronary artery disease. N Engl J Med 2007, 357(5):443-453.
71. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM et al: Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 2008, 40(2):161-169.
72. Ziegler A, Konig IR, Thompson JR: Biostatistical aspects of genome-wide association studies. Biom J 2008, 50(1):8-28.
73. Johnson JL, Green LW, Frankish CJ, MacLean DR, Stachenko S: A dissemination research agenda to strengthen health promotion and disease prevention. Can J Public Health 1996, 87 Suppl 2:S5-10.
74. Wolf SM, Lawrenz FP, Nelson CA, Kahn JP, Cho MK, Clayton EW, Fletcher JG, Georgieff MK, Hammerschmidt D, Hudson K et al: Managing incidental findings in human subjects research: analysis and recommendations. J Law Med Ethics 2008, 36(2):219-248, 211.
75. Ellerin BE, Schneider RJ, Stern A, Toniolo PG, Formenti SC: Ethical, legal, and social issues related to genomics and cancer research: the impending crisis. J Am Coll Radiol 2005, 2(11):919-926.
76. Van Hoyweghen I, Horstman K: European practices of genetic information and insurance: lessons for the Genetic Information Nondiscrimination Act. Jama 2008, 300(3):326-327.
77. Offit K: Genomic profiles for disease risk: predictive or premature? Jama 2008, 299(11):1353-1355.
78. Netzer C, Biller-Andorno N: Pharmacogenetic testing, informed consent and the problem of secondary information. Bioethics 2004, 18(4):344-360.
79. Eisenberg L: Is biology destiny? Is it all in our genes? J Psychiatr Pract 2002, 8(6):337-343.
80. Ozdemir V, Godard B: Evidence-based management of nutrigenomics expectations and ELSIs. Pharmacogenomics 2007, 8(8):1051-1062.
81. Levesque L, Ozdemir V, Gremmen B, Godard B: Integrating anticipated nutrigenomics bioscience applications with ethical aspects. Omics 2008, 12(1):1-16.
82. The Comprehensive R Archive Network. http://cran.r-project.org/.
83. MDR software. [http://www.epistasis.org].
84. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B: Microarray data mining with visual programming. Bioinformatics 2005, 21(3):396-398.
85. Witten IH, Frank E: Data mining: practical machine learning tools and techniques, Second edn. San Francisco: Morgan Kaufmann; 2005.
86. Afman L, Muller M: Nutrigenomics: from molecular nutrition to prevention of disease. J Am Diet Assoc 2006, 106(4):569-576.
87. McBride CM: Blazing a trail: a public health research agenda in genomics and chronic disease. Prev Chronic Dis 2005, 2(2):A04.
88. Lenzer J, Brownlee S: Knowing me, knowing you. Bmj 2008, 336(7649):858-860.
89. Smith SC, Jr.: Multiple risk factors for cardiovascular disease and diabetes mellitus. Am J Med 2007, 120(3 Suppl 1):S3-S11.
90. Astrup A: Healthy lifestyles in Europe: prevention of obesity and type II diabetes by diet and physical activity. Public Health Nutr 2001, 4(2B):499-515.

Supplemental data

Chapter 3 Color figures

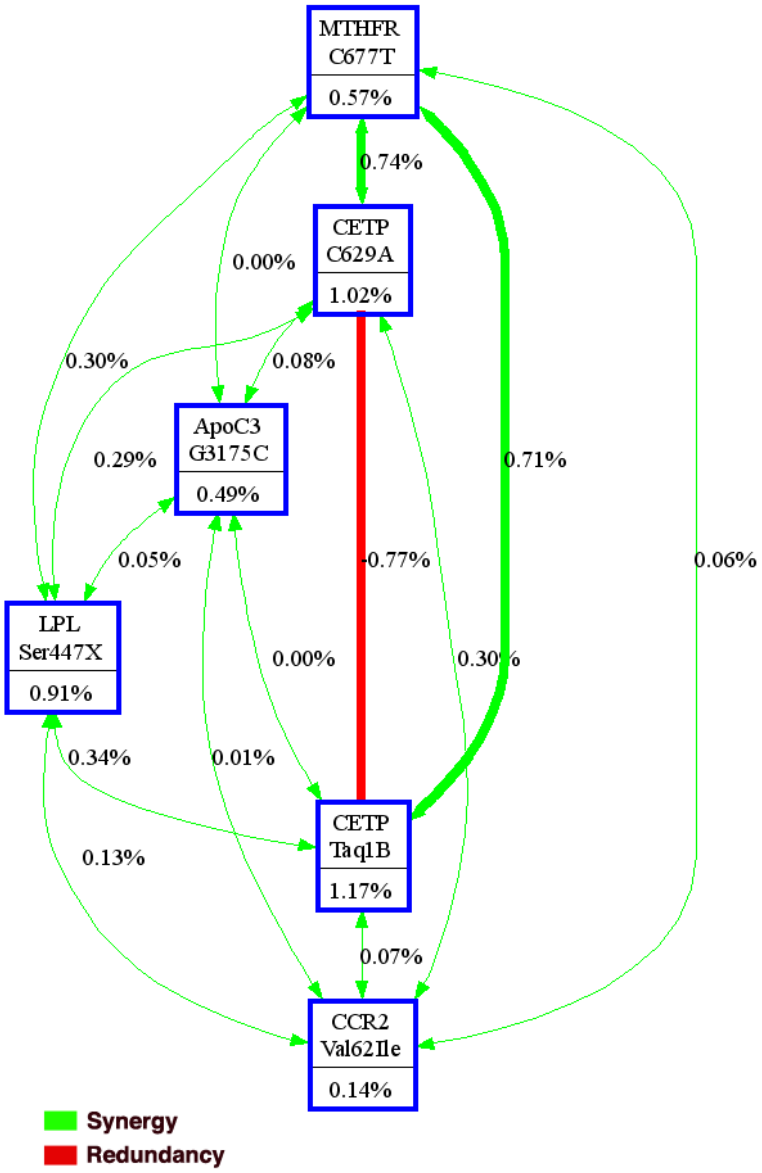


Figure 4: Entropy-based interaction graph. The percentages of entropy of HDL-cholesterol group explained by the different SNPs are shown in the boxes. The numbers by the arrows correspond to the percentages of entropy of HDL-cholesterol group explained by the two-way interactions between single nucleotide polymorphisms.

Chapter 3 Color figures (continued)

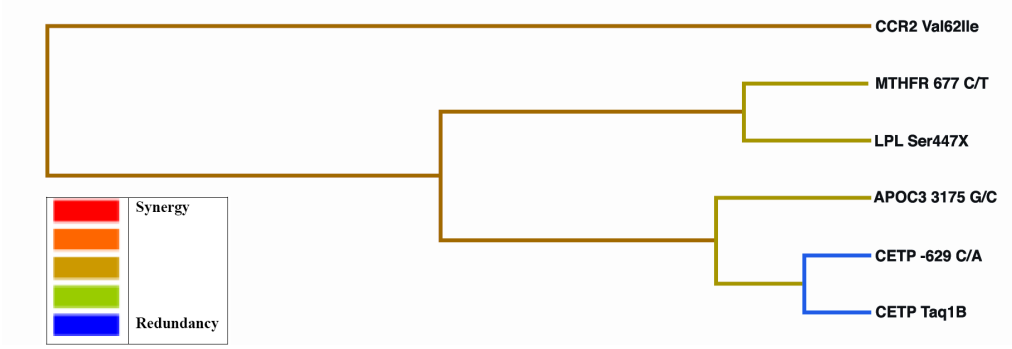


Figure 5: Interaction dendrogram. Stronger interactions between single nucleotide polymorphisms are visualized by depicting SNPs more closely together at the leaves of the tree (right side of the graph).

Chapter 5 Color figure

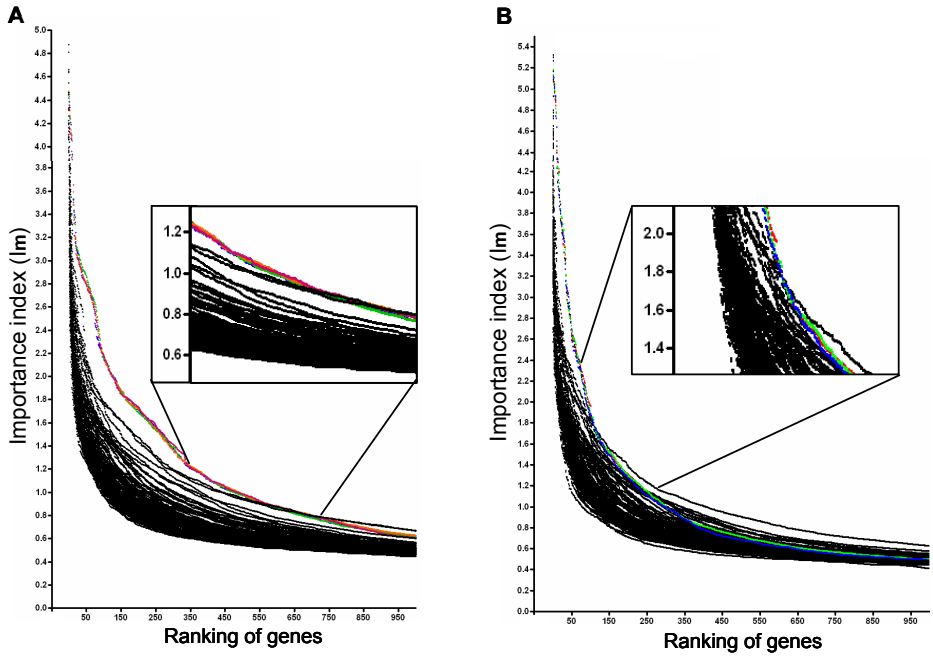


Figure 2A and 2B: Genes, of 100 random sets (black lines) and real sets with different seed values (colored lines), ranked by the I_m values. For colon (A) and cecum (B) datasets.

Chapter 3 Supplemental table: In the table information is provided for the SNPs genotyped in this study. For SNPs included in the analyses, minor allele frequencies and numbers of non-cases and cases for the genotypes are shown. For these SNPs p-values obtained by Fisher's Exact test are shown in the last column.

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
1	rs699	AGT met235thr (T/C)	C 0.42	T/T	183	188	0.65
				T/C	259	238	
				C/C	103	107	
1	rs6025	F5 arg506gln (G/A)	A 0.02	G/G	514	515	0.08
				G/A | A/A	31	18	
1	rs1800872	IL-10 -571 C/A	A 0.23	C/C	342	319	0.37
				C/A	176	178	
				A/A	27	36	
1	rs1801133	MTHFR ala222val (677 C/T)	T 0.32	C/C	235	277	0.01
				C/T	248	203	
				T/T	62	53	
1	rs5063	NPPA 664 G/A	A 0.04	G/G	508	489	0.42
				G/A | A/A	37	44	
1	rs5065	NPPA 2238 T/C	C 0.16	T/T	394	369	0.26
				T/C	139	144	
				C/C	12	20	
1	rs5361	SELE ser149arg (A/C) *	C 0.11	A/A	434	420	0.76
				A/C | C/C	111	113	
1	rs5355	SELE leu554phe (C/T)	T 0.04	C/C	501	500	0.24
				C/T | T/T	44	33	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
1	rs6131	SELP ser330asn (G/A)	A 0.18	G/G	378	360	0.48
				G/A	149	148	
				A/A	18	25	
1	rs6133	SELP val640leu (G/T)	T 0.12	G/G	417	419	0.66
				G/T	116	105	
				T/T	12	9	
1	rs1041163	VCAM-1 -1594 T/C	C 0.14	T/T	406	401	0.85
				T/C	126	122	
				C/C	13	10	
2	rs1367117	ApoB thr71ile (C/T)	T 0.32	C/C	264	246	0.13
				C/T	233	220	
				T/T	48	67	
2	rs5742904	ApoB arg3500gln (G/A)	minor allele frequency <0.01				
2	rs5742909	CTLA4 -318 C/T	T 0.07	C/C	488	460	0.11
				C/T | T/T	57	73	
2	rs231775	CTLA4 thr17ala (A/G)	G 0.38	A/A	205	204	0.35
				A/G	273	249	
				G/G	67	80	
2	rs1800587	IL-1A thr(-889)cys (T/C)	C 0.71	T/T	56	44	0.48
				T/C	198	204	
				C/C	291	285	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
2	rs16944	IL-1B -1418 C/T	T 0.34	C/C	231	223	0.48
				C/T	258	243	
				T/T	56	67	
2	rs1143634	IL-1B phe105phe (C/T)	T 0.24	C/C	321	310	0.82
				C/T	197	192	
				T/T	27	31	
3	rs5186	AGTR1 1166 A/C	C 0.30	A/A	263	267	0.84
				A/C	236	223	
				C/C	46	43	
3	rs1799864	CCR2 Val62Ile (G/A)	A 0.08	G/G	453	460	0.15
				G/A | A/A	92	73	
3	rs5742906	CCR3 Pro39Leu (C/T)	minor allele frequency <0.01				
3	rs333	CCR5 32-bp ins/del	del 0.11	ins/ins	433	429	0.70
				ins/del | del/del	112	104	
3	rs1799987	CCR5 -2459 G/A	A 0.55	G/G	107	106	0.61
				G/A	273	280	
				A/A	165	147	
3	rs2290608	IL-5ra -80 G/A	A 0.24	G/G	338	292	0.05
				G/A	176	207	
				A/A	31	34	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
3	rs1801282	PPARG pro12ala (C/G)	G 0.12	C/C	409	421	0.26
				C/G	125	105	
				G/G	11	7	
4	rs4961	ADD1 gly460trp (G/T)	T 0.20	G/G	350	356	0.55
				G/T	165	154	
				T/T	30	23	
4	rs1800790	FGB -455 G/A	A 0.19	G/G	363	357	0.75
				G/A	162	152	
				A/A	20	24	
4	rs7041	GC glu416asp (G/T)	T 0.45	G/G	160	167	0.62
				G/T	269	264	
				T/T	116	102	
4	rs4588	GC thr420lys (C/A)	A 0.28	C/C	289	274	0.23
				C/A	209	225	
				A/A	47	34	
5	rs1042713	ADRB2 arg16gly (A/G) *	G 0.60	A/A	78	92	0.35
				A/G	276	269	
				G/G	191	172	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
5	rs1042714	ADRB2 gln27glu (C/G) *	G 0.41	C/C	168	186	0.26
				C/G	287	273	
				G/G	90	74	
5	rs1800888	ADRB2 thr164ile (C/T)	T 0.01	C/C	532	524	0.52
				C/T | T/T	13	9	
5	rs2569190	CD14 -260 C/T	T 0.49	C/C	143	127	0.54
				C/T	275	286	
				T/T	127	120	
5	rs25882	CSF2 ile117thr (T/C)	C 0.22	T/T	326	345	0.07
				T/C	195	157	
				C/C	24	31	
5	rs2243250	IL-4 -590 C/T	T 0.17	C/C	359	376	0.18
				C/T	173	142	
				T/T	13	15	
5	rs2069885	IL-9 thr113met (C/T)	T 0.11	C/C	433	429	0.70
				C/T | T/T	112	104	
5	rs1295686	IL-13 4045 C/T	T 0.21	C/C	340	336	0.25
				C/T	185	167	
				T/T	20	30	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
5	rs1062535	ITGA2 873 G/A	A 0.39	G/G	203	199	0.98
				G/A	260	256	
				A/A	82	78	
5	rs730012	LTC4S -444 A/C	C 0.30	A/A	284	269	0.50
				A/C	206	218	
				C/C	55	46	
5	rs244656	TCF7 -1459 A/T	T 0.14	A/A	405	390	0.90
				A/T	129	131	
				T/T	11	12	
5	rs5742913	TCF7 Pro19Thr (C/T)	Hardy Weinberg Disequilibrium ($p<0.01$)				
6	rs1853021	LPA 93 C/T	T 0.15	C/C	372	396	0.04
				C/T	161	132	
				T/T	12	5	
6	rs1800769	LPA 121 G/A	A 0.15	G/G	398	392	0.97
				G/A	134	128	
				A/A	13	13	
6	rs1041981	LTA thr26asn (C/A)	A 0.34	C/C	227	241	0.50
				C/A	255	234	
				A/A	63	58	
6	rs909253	LTA intronA A/G	correlated with LTA thr26asn ($r>0.85$)				
6	rs1800750	TNF -376 G/A	Hardy Weinberg Disequilibrium ($p<0.01$)				

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
6	rs1800629	TNF -308 G/A*	A 0.17	G/G	379	362	0.60
				G/A	149	158	
				A/A	17	13	
6	rs673	TNF -244 G/A	minor allele frequency <0.01				
6	rs361525	TNF -238 G/A*	A 0.04	G/G	504	496	0.73
				G/A | A/A	41	37	
7	rs1800796	IL-6 -572 G/C	Hardy Weinberg Disequilibrium (p<0.01)				
7	rs1800795	IL-6 -174 G/C	C 0.38	G/G	229	200	0.30
				G/C	236	245	
				C/C	80	88	
7	rs1800779	NOS3 -922 A/G*	G 0.36	A/A	225	214	0.93
				A/G	251	250	
				G/G	69	69	
7	rs3918226	NOS3 -690 C/T	T 0.07	C/C	469	457	0.93
				C/T | T/T	76	76	
7	rs1799983	NOS3 glu298asp (G/T)*	T 0.31	G/G	262	251	0.94
				G/T	232	230	
				T/T	51	52	
7	rs1799768	PAI-1 -675 G5G4	del 0.53	G/G	118	132	0.35
				G/del	260	255	
				del/del	167	146	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
7	rs7242	PAI-1 11053 G/T	T 0.56	G/G	108	110	0.41
				G/T	270	243	
				T/T	167	180	
7	rs854560	PON1 met55leu (A/T)	T 0.64	A/A	68	67	0.04
				A/T	274	229	
				T/T	203	237	
7	rs662	PON1 gln192arg (A/G)	G 0.30	A/A	285	262	0.29
				A/G	199	219	
				G/G	61	52	
7	rs6954345	PON2 ser311cys (C/G)	G 0.25	C/C	317	295	0.56
				C/G	193	197	
				G/G	35	41	
8	rs4994	ADRB3 trp64arg (T/C)	C 0.08	T/T	470	454	0.91
				T/C	70	74	
				C/C	5	5	
8	rs1800590	LPL -93 T/G	Hardy Weinberg Disequilibrium ($p<0.01$)				
8	rs1801177	LPL asp9asn (G/A) *	A 0.02	G/G	531	511	0.18
				G/A | A/A	14	22	
8	rs268	LPL asn291ser (A/G) *	G 0.02	A/A	527	507	0.22
				A/G | G/G	18	26	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
8	rs328	LPL Ser447X (C/G) *	G 0.12	C/C	397	438	0.0008
				C/G	142	90	
				G/G	6	5	
9	rs17611	C5 ile802val (A/G)	G 0.59	A/A	95	85	0.81
				A/G	269	266	
				G/G	181	182	
10	rs1801157	SDF1 800 G/A	A 0.18	G/G	375	357	0.68
				G/A	148	157	
				A/A	22	19	
11	rs670	ApoAI -75 G/A	A 0.19	G/G	346	359	0.41
				G/A	180	157	
				A/A	19	17	
11	rs675	ApoA4 thr347ser (A/T)	T 0.19	A/A	350	353	0.68
				A/T	178	167	
				T/T	17	13	
11	rs5110	ApoA4 gln360his (G/T)	T 0.08	G/G	463	459	0.60
				G/T | T/T	82	74	
11	rs2542052	ApoC3 -641 C/A	Hardy Weinberg Disequilibrium ($p<0.01$)				
11	rs2854117	ApoC3 -482 C/T	T 0.28	C/C	292	264	0.34
				C/T	213	221	
				T/T	40	48	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
11	rs2854116	ApoC3 -455 T/C	C 0.37	T/T	219	203	0.42
				T/C	247	262	
				C/C	79	68	
11	rs4520	ApoC3 1100 C/T	T 0.29	C/C	281	261	0.54
				C/T	229	230	
				T/T	35	42	
11	rs5128	ApoC3 3175 G/C*	C 0.10	G/G	456	411	0.007
				G/C | C/C	89	122	
11	rs4225	ApoC3 3206 T/G	G 0.40	T/T	188	192	0.68
				T/G	275	255	
				G/G	82	86	
11	rs1799963	F2 20210 G/A	Hardy Weinberg Disequilibrium ($p<0.01$)				
11	rs569108	FCER1B glu237gly (A/G)	G 0.02	A/A	530	511	0.24
				A/G | G/G	15	22	
11	rs3025058	MMP3 -1171 A5A6	A 0.48	del/del	148	151	0.89
				del/A	267	260	
				A/A	130	122	
11	rs3741240	UGB 38 G/A	A 0.34	G/G	233	220	0.24
				G/A	262	247	
				A/A	50	66	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
12	rs5443	GNB3 825 C/T	T 0.30	C/C	268	262	0.95
				C/T	226	224	
				T/T	51	47	
12	rs5742912	SCNN1A trp493arg (T/C)	C 0.02	T/T	529	506	0.09
				T/C | C/C	16	27	
12	rs2228576	SCNN1A ala663thr (G/A)	A 0.34	G/G	227	241	0.43
				G/A	257	231	
				A/A	61	61	
12	rs2228570	VDR met1thr (T/C)	C 0.61	T/T	77	85	0.16
				T/C	278	241	
				C/C	190	207	
12	rs1544410	VDR intron8 G/A	A 0.40	G/G	187	203	0.41
				G/A	266	241	
				A/A	92	89	
13	rs5742910	F7 -323 10-bp del/ins	ins 0.11	del/del	436	414	0.63
				del/ins	104	114	
				ins/ins	5	5	
13	rs6046	F7 arg353gln (G/A)	A 0.09	G/G	457	429	0.15
				G/A | A/A	88	104	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
15	rs1800588	LIPC -480 C/T	T 0.12	C/C	317	331	0.40
				C/T	202	181	
				T/T	26	21	
16	rs1800776	CETP -631 C/A	A 0.08	C/C	476	440	0.03
				C/A | A/A	69	93	
16	rs1800775	CETP -629 C/A	A 0.46	C/C	139	183	0.0005
				C/A	268	258	
				A/A	138	92	
16	rs708272	CETP Taq1B (G/A)*	A 0.39	G/G	172	217	0.0002
				G/A	270	257	
				A/A	103	59	
16	rs5882	CETP ile405val (A/G)*	G 0.31	A/A	252	267	0.19
				A/G	227	218	
				G/G	66	48	
16	rs2303790	CETP asp442gly (A/G)	minor allele frequency <0.01				
16	-	CETP intron14 1 G/A	Hardy Weinberg Disequilibrium (p<0.01)				
16	-	CETP intron14 3 insT	Hardy Weinberg Disequilibrium (p<0.01)				
16	rs1805010	IL4R ile75val (A/G)	G 0.46	A/A	136	152	0.21
				A/G	294	288	
				G/G	115	93	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
16	rs1805015	IL-4R ser503pro (T/C)	C 0.17	T/T	375	380	0.64
				T/C	154	137	
				C/C	16	16	
16	rs1801275	IL-4R gln576arg (A/G)	G 0.22	A/A	327	336	0.26
				A/G	193	166	
				G/G	25	31	
17	rs1799752	ACE intron16 ins/del	Hardy Weinberg Disequilibrium ($p<0.01$)				
17	rs5918	ITGB3 leu33pro (T/C)	C 0.16	T/T	381	386	0.65
				T/C	147	132	
				C/C	17	15	
17	rs1137933	NOS2A asp346asp (C/T)	T 0.22	C/C	349	319	0.01
				C/T	162	195	
				T/T	34	19	
17	rs4795895	SCYA11 -1328 G/A	A 0.17	G/G	381	374	0.97
				G/A	146	143	
				A/A	18	16	
17	rs3744508	SCYA11 ala23thr (G/A)	A 0.18	G/G	363	376	0.38
				G/A	160	138	
				A/A	22	19	

Chapter 3 Supplemental table (continued)

Chromosome	rs number	SNP	minor allele frequency	genotype	non-cases	cases	Fisher´s Exact test p-value
19	rs429358	ApoE cys112arg (T/C)*	C 0.14	T/T	408	390	0.81
				T/C	124	129	
				C/C	13	14	
19	rs7412	ApoE arg158cys (C/T)*	T 0.07	C/C	468	463	0.66
				C/T | T/T	77	70	
19	rs2230199	C3 arg102gly (C/G)	G 0.19	C/C	355	358	0.07
				C/G	174	147	
				G/G	16	28	
19	rs5491	ICAM-1 Lys56Met (A/T)	minor allele frequency <0.01				
19	rs1799969	ICAM-1 gly241arg (G/A)*	A 0.12	G/G	423	407	0.66
				G/A	111	118	
				A/A	11	8	
19	rs5742911	LDLR Ncol +/–	Hardy Weinberg Disequilibrium (p<0.01)				
19	rs1800469	TGFB1 -509 C/T	T 0.26	C/C	301	290	0.58
				C/T	205	212	
				T/T	39	31	
21	rs5742905	CBS ile278thr 68-bp ins	large numbers of missing values				

* Genotyped by two different methods

Chapter 7 Color figures

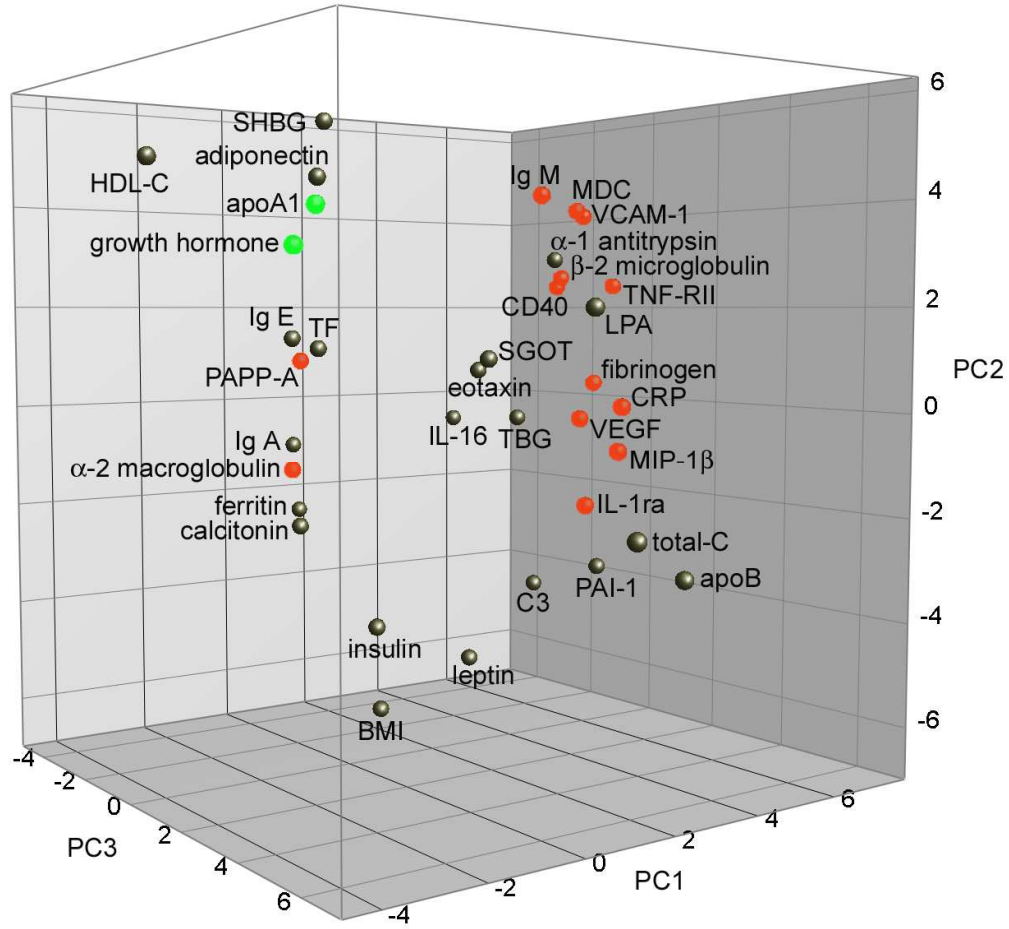


Figure 3A: PCA loading plot including selected proteins and the intermediate endpoints BMI, HDL-C and total-C. In the PCA loading plot the proteins are plotted based on their loadings on the first three principal components. The further proteins are located from the origin, the stronger their contribution in explaining the variance. As most variance is explained by the first component, proteins that have high loadings on the first component contribute most. Proteins and intermediate endpoints grouped together show positive relationships with each other, whereas proteins and intermediate endpoints that are located opposite to each other (as seen from the origin) show negative relationships with each other. Proteins in red and green were respectively positively and negatively associated with CHD mortality.

Chapter 7 Color figures (continued)

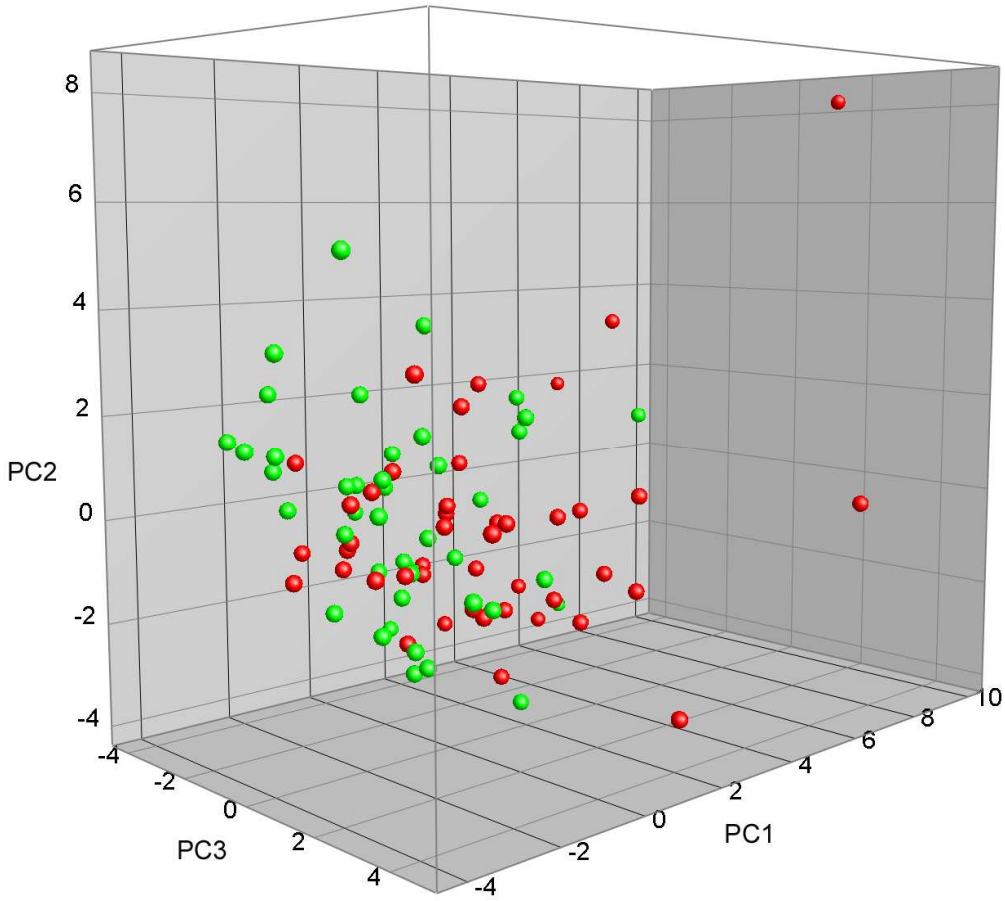


Figure 3B: PCA score plot corresponding to the loading plot. In the PCA score plot the individuals are plotted based on their scores on the first three principal components. Red dots indicate cases of CHD mortality, green dots indicate controls. Observations that are located in the direction of certain proteins score high on these proteins, whereas observations that are located in the opposite direction of certain proteins score low on these proteins (see also figure 3A).

Chapter 7 Supplemental table 1: Univariate associations of 83 proteins with CHD mortality.

Protein	Unit	CHD mortality cases	controls	p-value
1 α-1 Antitrypsin	mg/ml	1.71 [1.45-2.04]	1.60 [1.34-1.81]	0.19
2 α-2_Macroglobulin	mg/ml	6.24 [5.43-10.2]	6.16 [4.93-8.66]	0.61
3 α Fetoprotein	ng/ml	2.12 [1.84-2.57]	2.15 [1.79-2.70]	0.83
4 Adiponectin	µg/ml	3.29 ± 1.36	3.46 ± 1.17	0.55
5 ApoA1	mg/ml	0.33 [0.28-0.38]	0.36 [0.32-0.43]	0.02*
6 ApoB	mg/ml	1.41 ± 0.29	1.24 ± 0.28	0.007**
7 ApoC3	µg/ml	141 ± 39.9	156 ± 44.8	0.10
8 ApoH	µg/ml	180 [162-212]	177 [159-203]	0.75
9 β-2 Microglobulin	µg/ml	1.84 [1.66-2.25]	1.62 [1.47-1.95]	0.009**
10 BDNF	ng/ml	3.65 [1.41-7.96]	2.94 [1.37-6.44]	0.69
11 Complement 3	mg/ml	1.33 ± 0.21	1.26 ± 0.29	0.25
12 CD40	ng/ml	0.33 [0.26-0.37]	0.31 [0.26-0.36]	0.24
13 CD40 Ligand	ng/ml	0.07 [0.05-0.18]	0.08 [0.04-0.14]	0.66
14 CRP	µg/ml	3.47 [1.78-7.05]	1.86 [0.86-4.04]	0.01*
15 Calcitonin	pg/ml	2.46 [0.60-3.57]	1.91 [0.60-3.77]	0.91
16 CA 125	U/ml	6.83 ± 3.35	7.11 ± 3.66	0.71
17 CA 19-9	U/ml	1.62 [1.02-2.60]	1.52 [0.79-4.15]	0.89
18 CEA	ng/ml	1.54 [1.03-2.56]	1.47 [1.19-1.75]	0.51
19 CK-MB	ng/ml	0.26 [0.18-0.32]	0.24 [0.19-0.32]	0.78
20 EGF	pg/ml	1.81 [0.74-7.59]	0.74 [0.74-5.55]	0.30
21 ENA-78	ng/ml	0.77 [0.48-1.41]	0.63 [0.38-1.08]	0.54
22 ENRAGE	ng/ml	6.80 [4.00-14.4]	7.04 [4.36-13.4]	0.92
23 Endothelin-1	pg/ml	12.4 [9.07-15.3]	11.5 [8.62-15.3]	0.69
24 Eotaxin	pg/ml	93.5 [79.2-114]	88.1 [64.8-115]	0.27
25 Factor VII	ng/ml	364 ± 154	406 ± 125	0.16
26 FABP	ng/ml	0.70 [0.53-0.98]	0.65 [0.52-0.95]	0.68
27 Ferritin	ng/ml	232 [127-280]	238 [134-352]	0.35
28 Fibrinogen	mg/ml	3.67 [2.91-4.87]	3.27 [2.77-4.17]	0.12
29 GM-CSF	pg/ml	53.3 [43.7-61.4]	53.0 [44.0-61.4]	0.77
30 G-CSF	pg/ml	4.65 [0.50-5.85]	4.76 [0.50-7.54]	0.61
31 Growth Hormone	ng/ml	0.12 [0.10-0.19]	0.16 [0.10-0.29]	0.22
32 Haptoglobin	mg/ml	1.85 [1.05-2.47]	1.27 [0.93-1.89]	0.14
33 ICAM-1	ng/ml	122 ± 31.6	115 ± 30.8	0.28
34 IL-2	pg/ml	22.2 ± 8.33	22.7 ± 6.27	0.76
35 IL-3	ng/ml	0.22 ± 0.11	0.25 ± 0.12	0.25
36 IL-4	pg/ml	45.0 ± 16.7	45.2 ± 13.6	0.95
37 IL-5	pg/ml	12.4 ± 5.45	12.7 ± 4.53	0.79

Chapter 7 Supplemental table 1 (continued)

Protein	Unit	CHD mortality cases	controls	p-value
38 IL-7	pg/ml	115 ± 32.6	112 ± 23.7	0.70
39 IL-8	pg/ml	17.2 [12.8-18.8]	15.1 [12.8-17.1]	0.21
40 IL-10	pg/ml	15.7 [12.2-18.3]	15.0 [12.2-16.5]	0.33
41 IL-13	pg/ml	41.5 ± 16.8	42.5 ± 13.9	0.77
42 IL-15	ng/ml	1.03 [0.88-1.13]	1.04 [0.93-1.07]	0.63
43 IL-16	pg/ml	593 [523-668]	568 [487-650]	0.41
44 IL-18	pg/ml	270 [218-321]	256 [223-293]	0.57
45 IL-12p40	ng/ml	0.85 ± 0.32	0.85 ± 0.21	0.90
46 IL-12p70	pg/ml	328 ± 70.7	346 ± 64.3	0.22
47 IL-1ra	pg/ml	38.3 [26.3-64.4]	31.2 [20.0-46.8]	0.10
48 Ig A	mg/ml	1.88 ± 0.74	1.92 ± 0.67	0.79
49 Ig E	ng/ml	53.2 [28.3-127]	42.6 [31.6-119]	0.78
50 Ig M	mg/ml	1.03 [0.71-1.30]	0.86 [0.66-1.12]	0.12
51 Insulin	µlU/ml	6.32 [2.10-14.2]	5.89 [2.23-11.2]	0.84
52 LPA	µg/ml	55.9 [31.9-158]	39.4 [31.2-66.9]	0.10
53 Leptin	ng/ml	2.69 [1.77-4.50]	3.03 [1.70-4.36]	0.55
54 Lymphotactin	ng/ml	0.35 [0.29-0.39]	0.31 [0.27-0.41]	0.50
55 MCP-1	pg/ml	138 [113-160]	139 [123-155]	0.88
56 MDC	pg/ml	357[287-465]	346 [301-419]	0.58
57 MIP-1α	pg/ml	39.3 ± 11.3	36.9 ± 8.26	0.25
58 MIP-1β	pg/ml	168 [128-201]	153 [128-182]	0.21
59 MMP-2	ng/ml	774 [635-1060]	779 [648-1020]	1.00
60 MMP-3	ng/ml	0.16 [0.11-0.29]	0.19 [0.13-0.27]	0.54
61 MMP-9	ng/ml	208 [99.1-287]	146 [96.2-275]	0.32
62 Myeloperoxidase	ng/ml	50.3 [27.4-107]	39.8 [24.1-95.8]	0.45
63 Myoglobin	ng/ml	11.1 [8.04-14.9]	10.3 [7.64-13.7]	0.48
64 PAI-1	ng/ml	95.7 [76.1-115]	81.5 [59.3-101]	0.05*
65 PAP	ng/ml	0.12 [0.09-0.15]	0.12 [0.09-0.14]	0.80
66 PAPP-A	mlU/ml	0.05 [0.04-0.07]	0.04 [0.03-0.05]	0.04*
67 PSA, Free	ng/ml	0.40 [0.27-0.62]	0.45 [0.30-0.64]	0.61
68 RANTES	ng/ml	19.3 [9.63-28.1]	16.7 [9.82-26.9]	0.72
69 SGOT	µg/ml	11.4 ± 2.87	11.1 ± 2.80	0.63
70 SHBG	nmol/l	28.0 [23.8-36.2]	31.6 [22.0-42.3]	0.48
71 Serum Amyloid P	µg/ml	39.1 ± 10.2	39.1 ± 9.56	0.99
72 Stem Cell Factor	pg/ml	169 [117-229]	131 [114-182]	0.19
73 TBG	µg/ml	46.8 ± 10.3	48.1 ± 13.9	0.64
74 TIMP-1	ng/ml	80.4 [64.3-97.5]	76.4 [66.5-88.7]	0.73
75 TNF-α	pg/ml	5.67 ± 1.58	5.31 ± 1.38	0.26

Chapter 7 Supplemental table 1 (continued)

Protein	Unit	CHD mortality cases	controls	p-value
76 TNF-β	pg/ml	45.2 ± 16.8	44.6 ± 12.0	0.84
77 TNF RII	ng/ml	3.34 [2.65-4.14]	3.03 [2.54-3.39]	0.08
78 TSH	µlU/ml	1.07 [0.76-1.49]	1.01 [0.75-1.36]	0.60
79 Thrombopoietin	ng/ml	2.21 ± 0.59	2.12 ± 0.68	0.49
80 Tissue Factor	ng/ml	1.12 ± 0.19	1.10 ± 0.25	0.73
81 VCAM-1	ng/ml	398 [338-463]	349 [319-412]	0.03*
82 VEGF	pg/ml	264 [232-327]	229 [213-260]	0.004**
83 vWF	µg/ml	4.77 [1.92-11.3]	3.15 [2.00-5.69]	0.38

Numbers of the proteins correspond with the numbers of the proteins in figures 1A-D. Results are presented as mean ± SD for proteins that were normally distributed or as median [interquartile range] for proteins with skewed distributions.
* Significant at the 0.05 level
** Significant at the 0.01 level

Chapter 7 Supplemental table 2: Univariate associations of 83 proteins with BMI, HDL-C and total-C.

Protein	BMI		HDL-C		Total-C	
	r	p-value	r	p-value	r	p-value
1 α-1 Antitrypsin	-0.17	0.11	-0.17	0.11	0.03	0.81
2 α-2_Macroglobulin	0.005	0.96	-0.02	0.89	0.04	0.68
3 α Fetoprotein	-0.06	0.61	0.06	0.60	0.24	0.03*
4 Adiponectin	-0.26	0.01*	0.35	0.0007**	-0.02	0.88
5 ApoA1	-0.22	0.04*	0.50	<0.0001**	0.10	0.38
6 ApoB	0.18	0.10	-0.37	0.0004**	0.87	<0.0001**
7 ApoC3	0.02	0.87	0.19	0.07	0.08	0.46
8 ApoH	-0.05	0.61	-0.20	0.07	0.02	0.86
9 β-2 Microglobulin	0.08	0.47	-0.30	0.005**	0.003	0.98
10 BDNF	-0.06	0.57	-0.06	0.55	0.13	0.23
11 Complement 3	0.40	0.0001**	-0.46	<0.0001**	0.12	0.28
12 CD40	0.10	0.35	-0.22	0.04*	0.02	0.88
13 CD40 Ligand	-0.05	0.62	-0.10	0.35	0.09	0.42
14 CRP	0.18	0.10	-0.39	0.0002**	0.16	0.13
15 Calcitonin	0.15	0.15	-0.15	0.15	-0.08	0.44
16 CA 125	0.02	0.84	0.00	0.98	-0.09	0.39
17 CA 19-9	0.02	0.82	0.11	0.32	-0.19	0.08
18 CEA	-0.10	0.37	0.06	0.60	0.10	0.37
19 CK-MB	0.13	0.23	0.08	0.45	-0.13	0.22
20 EGF	-0.14	0.19	-0.04	0.71	0.03	0.76
21 ENA-78	0.09	0.43	-0.11	0.31	-0.04	0.70
22 ENRAGE	0.02	0.85	-0.08	0.46	0.16	0.14
23 Endothelin-1	-0.11	0.30	0.13	0.24	0.07	0.52
24 Eotaxin	0.05	0.65	-0.04	0.71	-0.01	0.91
25 Factor VII	0.006	0.95	-0.04	0.69	0.23	0.03*
26 FABP	0.07	0.51	0.03	0.80	0.01	0.93
27 Ferritin	0.33	0.002**	-0.08	0.47	-0.03	0.75
28 Fibrinogen	0.07	0.53	-0.26	0.01*	0.08	0.44
29 GM-CSF	0.006	0.95	0.08	0.45	-0.004	0.97
30 G-CSF	0.09	0.41	-0.09	0.40	0.14	0.18
31 Growth Hormone	-0.31	0.003**	0.23	0.03*	0.001	0.99
32 Haptoglobin	-0.06	0.58	-0.21	0.05*	0.16	0.14
33 ICAM-1	0.11	0.29	-0.30	0.005**	-0.08	0.45
34 IL-2	-0.07	0.52	0.01	0.93	-0.06	0.57
35 IL-3	0.03	0.77	-0.10	0.34	0.009	0.94
36 IL-4	0.01	0.93	0.01	0.91	-0.06	0.57
37 IL-5	-0.06	0.56	0.21	0.05*	-0.12	0.26

Chapter 7 Supplemental table 2 (continued)

Protein	BMI		HDL-C		Total-C	
	r	p-value	r	p-value	r	p-value
38 IL-7	0.05	0.65	0.04	0.73	-0.05	0.66
39 IL-8	0.08	0.48	-0.21	0.05*	0.01	0.92
40 IL-10	-0.001	0.99	-0.03	0.75	-0.09	0.42
41 IL-13	-0.14	0.18	0.17	0.12	0.08	0.47
42 IL-15	-0.05	0.64	-0.15	0.17	-0.03	0.75
43 IL-16	0.13	0.24	-0.29	0.006**	-0.10	0.37
44 IL-18	0.04	0.73	-0.16	0.13	-0.06	0.57
45 IL-12p40	-0.004	0.97	-0.01	0.95	-0.10	0.37
46 IL-12p70	0.16	0.13	-0.04	0.69	0.04	0.73
47 IL-1ra	0.31	0.004**	-0.31	0.003**	0.05	0.67
48 Ig A	0.10	0.36	-0.07	0.52	-0.24	0.03*
49 Ig E	-0.04	0.73	0.19	0.08	-0.01	0.91
50 Ig M	0.01	0.92	0.05	0.65	0.006	0.96
51 Insulin	0.44	<0.0001**	-0.36	0.0005**	-0.03	0.75
52 LPA	-0.16	0.14	-0.05	0.66	0.18	0.10
53 Leptin	0.72	<0.0001**	-0.28	0.008**	0.22	0.04*
54 Lymphotactin	-0.04	0.73	0.00	0.97	-0.03	0.79
55 MCP-1	0.07	0.51	-0.05	0.66	0.41	<0.0001**
56 MDC	-0.08	0.49	-0.13	0.24	-0.04	0.68
57 MIP-1α	-0.20	0.06	-0.06	0.56	-0.08	0.44
58 MIP-1β	0.16	0.14	-0.19	0.07	0.18	0.10
59 MMP-2	-0.13	0.23	0.04	0.71	-0.04	0.72
60 MMP-3	-0.22	0.04*	0.15	0.15	0.07	0.54
61 MMP-9	0.05	0.66	-0.06	0.58	0.03	0.75
62 Myeloperoxidase	0.11	0.30	-0.20	0.06	-0.04	0.69
63 Myoglobin	0.09	0.42	-0.05	0.64	-0.02	0.88
64 PAI-1	0.21	0.04*	-0.45	<0.0001**	0.22	0.04*
65 PAP	-0.15	0.17	-0.02	0.87	-0.03	0.81
66 PAPP-A	0.17	0.11	-0.04	0.73	-0.18	0.09
67 PSA, Free	-0.12	0.25	0.13	0.23	0.06	0.60
68 RANTES	0.02	0.87	-0.05	0.63	0.01	0.92
69 SGOT	0.04	0.73	0.03	0.81	0.16	0.15
70 SHBG	-0.45	<0.0001**	0.18	0.09	-0.03	0.75
71 Serum Amyloid P	0.18	0.10	-0.24	0.02*	0.19	0.07
72 Stem Cell Factor	0.14	0.19	-0.20	0.06	-0.02	0.88
73 TBG	0.12	0.26	-0.42	<0.0001**	0.11	0.29
74 TIMP-1	0.16	0.15	-0.15	0.16	0.008	0.94
75 TNF-α	-0.10	0.34	-0.07	0.51	-0.11	0.31

Chapter 7 Supplemental table 2 (continued)

Protein	BMI		HDL-C		Total-C	
	r	p-value	r	p-value	r	p-value
76 TNF-β	-0.03	0.78	0.07	0.55	-0.08	0.46
77 TNF RII	0.03	0.78	-0.29	0.007**	-0.007	0.95
78 TSH	-0.03	0.77	-0.18	0.10	0.03	0.75
79 Thrombopoietin	-0.06	0.59	-0.04	0.71	0.02	0.84
80 Tissue Factor	-0.19	0.08	0.08	0.46	-0.08	0.45
81 VCAM-1	0.02	0.83	-0.19	0.08	-0.06	0.59
82 VEGF	0.13	0.22	-0.29	0.007**	0.12	0.26
83 vWF	0.004	0.97	-0.15	0.17	0.02	0.88

Numbers of the proteins correspond with the numbers of the proteins in figures 1A-D. Results are presented as mean ± SD for proteins that were normally distributed or as median [interquartile range] for proteins with skewed distributions.

* Significant at the 0.05 level

** Significant at the 0.01 level

Summary

Summary

In the field of nutrigenomics large numbers of variables can nowadays be obtained due to the development in technology. This provides the opportunity for researchers not only to study the individual effects of genes and proteins, but also the combined effects of groups of genes and/or proteins in relation to complex diseases. In complex diseases it is assumed that many factors play a role, whereby each variable by itself will only have a small to moderate effect. Furthermore, factors likely contribute to complex diseases by intricate and ubiquitous interactions. The large numbers of variables obtained in moderate to small sample sizes and the complexity by which these variables can relate to the outcome of interest has led to challenges in the statistical analysis of nutrigenomic datasets. In this thesis approaches to analyze large nutrigenomic datasets are investigated, including the analysis of genetic, transcriptomic and proteomic data. These approaches are incorporated within a general framework for nutrigenomic data analysis with the objective to obtain insight in biological mechanisms involved in the development of complex diseases. Applications of these approaches to real data to obtain biologically relevant information or to obtain models for diagnostic or prognostic purposes are also presented.

An overview of methods to analyze SNP data is provided in chapter 2, in which the strengths and weaknesses of both traditional statistical methods and non-parametric methods are discussed. Traditional statistical methods are not suitable to analyze large numbers of single nucleotide polymorphisms (SNPs), because these methods can only handle a limited number of variables in moderate to small sample sizes. On the other hand, non-parametric methods are currently available that are able to overcome the statistical challenges in the analyses of large numbers of SNPs. As these methods have different selection features, the main conclusion of chapter 2 is that a combination of several non-parametric methods seems to be a useful strategy to analyze SNPs in relation to complex diseases. To investigate this conjecture, we carried out a study in which we applied different non-parametric methods to a real dataset to compare their prioritization and selection of SNPs, as is described in chapter 3. In this study we analyzed the association of 93 SNPs with plasma HDL-cholesterol levels. The results show that applying a combination of different methods to the same dataset has advantages compared to applying only one method. After selection of a subset of SNPs, we applied interaction graphs for statistical interpretation. The interaction entropy graph appeared to be a useful tool to visualize whether SNPs contribute to the outcome of interest by their main and/or interaction effect. For statistical interpretation we additionally applied logistic regression analysis, which can also be useful at this stage if the number of observations is sufficiently large relative to the number of selected SNPs. In chapter 4 we applied both the interaction entropy graph and logistic regression analysis for statistical interpretation of a biological model of weight regulation. In this study we identified sex-specific associations of CNTF, IL6 and UCP2 polymorphisms with weight gain. These associations were shown to be independent of leptin.

In general, sample sizes are smaller in microarray studies compared to genetic epidemiological studies. Therefore the statistical analysis of microarray data is even more challenging. In nutrigenomic datasets the effects of genes are likely to be small, but in combination these genes may have an important effect. Therefore, to detect these combined

effects, it is necessary to take in the selection of genes interactions into account. In that respect, the features of random forests (RF) make it an attractive method for the analysis of microarray data: besides the ability to handle large numbers of variables in small sample sizes, it provides an importance index for each gene in which all possible interactions with other genes are taken into account. In chapter 5 we developed a framework for the analysis of microarray data, taking both GO-annotated and non-annotated genes into account. We applied RF to two real microarray datasets to show its advantage in the selection of genes. For statistical interpretation, genes selected by RF were subsequently analyzed by Self-Organizing Maps (SOM) to cluster genes with similar gene expression profiles. It appeared that we identified clusters consisting of genes that were only important in interaction with other genes. Genes within these clusters were found to belong to the same biological process and therefore have biological meaning. This indicates the importance of taking interactions in the selection of genes into account. Thus, in this study we showed that applying RF in combination with a cluster method helped in retrieving biologically relevant information from microarray data.

In proteomics, mass spectrometry data can be analyzed to detect biomarkers for diagnostic purposes. As part of a classification contest to best classify and predict breast cancer cases and controls, we applied a three-step approach to analyze mass spectrometry data to come to a discrimination rule (see chapter 6). Peptide masses were prioritized by RF. As the top-ranked variables consisted of highly correlated variables, new variables were created to group these highly correlated variables. These newly created variables were finally included in a model to predict breast cancer cases and controls, yielding a good classifying and interpretable model. The sensitivity and specificity of this model was 86.8% and 85.7%, respectively.

To overcome the problem of high-dimensional data, methods exist that reduce the dimensionality of the data by creating a smaller set of variables. For example, partial least squares (PLS) reduces the dimensionality of the data by constructing latent components. In chapter 7 we applied PLS to study the association of 83 plasma proteins with coronary heart disease (CHD) mortality and intermediate endpoints involved in the etiology of CHD, namely body mass index (BMI), HDL-cholesterol (HDL-C) and total cholesterol (total-C). In this way we identified a set of 15 proteins which predicted 65% of CHD mortality, and sets of proteins associated with BMI, HDL-C and total-C. Subsequently, we analyzed identified sets of proteins together with intermediate endpoints by applying principal components analysis (PCA). This provided a comprehensive view of the relationships between identified proteins, intermediate endpoints and CHD mortality, showing that proteins involved in inflammation explained most of the variance, followed by proteins involved in metabolism and proteins related to total-C.

In this thesis approaches to analyze large nutrigenomic datasets have been discussed that overcome the statistical challenges present in this type of data. Applying these approaches to real nutrigenomic datasets, we were able to obtain clear interpretable statistical models and retrieve biologically meaningful information from the data. The results provided insight in the genes and proteins involved in intermediate endpoints and in complex diseases. The next step in nutrigenomic research is to combine data from different biological levels in the systems biology approach to generate more complete biological hypotheses regarding the development of complex diseases in humans. Hopefully, nutrigenomic research will in the

near future also lead to detection of more biomarkers with clinical relevance. The translation of genomic research into public health benefits has been limited up till now, but new advances in genomic research are promising, and may lead to applications that improve public health.

Samenvatting

Samenvatting

In het vakgebied van nutrigenomics kunnen door de huidige ontwikkeling van de technologie grote aantallen variabelen worden verkregen. Dit geeft onderzoekers de mogelijkheid om niet alleen individuele effecten van genen en eiwitten, maar ook gecombineerde effecten van groepen van genen en/of eiwitten in relatie tot complexe ziektes te bestuderen. Veel factoren spelen een rol bij complexe ziektes door middel van vele interacties, waarbij iedere factor afzonderlijk een klein effect zal hebben. De grote aantallen variabelen verkregen in relatief kleine steekproeven en de complexiteit waardoor deze variabelen gerelateerd kunnen zijn aan de ziekte, hebben geleid tot uitdagingen voor de statistische analyses van nutrigenomic datasets. In deze thesis zijn manieren onderzocht voor het analyseren van grote nutrigenomic datasets, inclusief het analyseren van genetische, transcriptomic and proteomic data. Deze aanpak is ondergebracht in een algemeen kader voor het analyseren van nutrigenomic data.

Een overzicht van methodes voor het analyseren van single nucleotide polymorphism (SNP) data is beschreven in hoofdstuk 2. In dit hoofdstuk worden de sterke en zwakke kanten van zowel traditionele statistische methodes als non-parametrische methodes besproken. Traditionele methodes zijn niet geschikt voor het analyseren van grote aantallen SNPs, aangezien het met deze methodes alleen mogelijk is om een beperkte hoeveelheid variabelen te analyseren bij relatief kleine steekproeven. Daarentegen zijn non-parametrische methodes in staat om te gaan met de statistische moeilijkheden die zich voordoen in het analyseren van grote aantallen SNPs. Deze methodes hebben verschillende kenmerken in het selecteren van de belangrijkste SNPs. Daarom is de conclusie van hoofdstuk 2 dat een combinatie van verscheidene non-parametrische methodes een bruikbare strategie lijkt te zijn voor het analyseren van grote aantallen SNPs in relatie tot complexe ziektes. Om deze stelling te onderzoeken hebben we een studie uitgevoerd waarin we verschillende non-parametrische methodes hebben toegepast op een echte dataset om de rangschikking en selectie van SNPs te vergelijken. Dit is beschreven in hoofdstuk 3. In deze studie hebben we de associatie van 93 SNPs met plasma HDL-cholesterol niveaus geanalyseerd. De resultaten laten zien dat het toepassen van een combinatie van meerdere methodes op dezelfde dataset voordelen heeft boven het toepassen van slechts één methode. Na het selecteren van de belangrijkste SNPs hebben we interactiegrafieken toegepast om statistisch te interpreteren of SNPs op zich al een effect hebben of een effect hebben in interactie met andere SNPs. Daarnaast hebben we logistische regressie-analyse toegepast. Het gebruik van de interactie-entropy grafiek in combinatie met logistische regressie-analyse bleek een goede aanpak te zijn voor het statistisch interpreteren van de bijdrage van SNPs aan de uitkomst, zij het door middel van hoofd- en/of interactie-effecten. In hoofdstuk 4 hebben we zowel de interactie-entropy grafiek alsook logistische regressie-analyse toegepast om een biologisch model betreffende gewichtsregulatie statistisch te interpreteren. In deze studie hebben we geslachtsspecifieke associaties van CNTF, IL6 and UCP2 polymorfismes met gewichtstoename geïdentificeerd.

Over het algemeen zijn in transcriptomic studies de steekproeven kleiner dan in genetisch epidemiologische studies, wat het statistisch analyseren van microarray-data nog moeilijker maakt. In nutrigenomic datasets zijn de effecten van genen naar verwachting klein, maar in combinatie kunnen deze genen een belangrijk effect hebben. Het is daarom noodzakelijk

om bij de selectie van genen rekening te houden met interacties. Wat dat betreft is random forests (RF) een aantrekkelijke methode voor het analyseren van microarray-data, aangezien deze om kan gaan met grote aantallen variabelen bij kleine steekproeven, maar ook omdat het een waarde van belangrijkheid toekent aan elk gen waarbij alle mogelijke interacties met andere genen zijn inbegrepen. In hoofdstuk 5 hebben we een kader ontwikkeld voor het analyseren van microarray-data, daarbij rekening houdend met alle genen aanwezig in de dataset. We hebben RF op twee echte microarray-datasets toegepast om de voordelen van deze methode in het selecteren van genen te laten zien. Voor de geselecteerde genen hebben we onderzocht of deze genen op zich een effect hebben of in interactie met andere genen. Voor de statistische interpretatie hebben we Self-Organizing Maps (SOM) toegepast om de genen met vergelijkbare genexpressie-profielen te groeperen in clusters. Met SOM hebben we clusters geïdentificeerd die bestonden uit genen die alleen belangrijk waren in interactie met andere genen. Genen binnen deze clusters behoorden tot hetzelfde biologische proces. Dit geeft aan dat het belangrijk is om bij de selectie van genen rekening te houden met interacties. Dus in deze studie tonen wij aan dat het toepassen van RF in combinatie met een clustermethode (bijvoorbeeld SOM) helpt in het verkrijgen van biologisch relevante informatie uit de microarray-dataset.

In proteomics kunnen massa-spectrometrie-data worden geanalyseerd om biomarkers te detecteren voor diagnostische en prognostische doeleinden. Als onderdeel van een competitie – die tot doel had om een aanpak te vinden die het beste onderscheid kan maken (discrimineren) tussen patiënten met borstkanker en controlepersonen – hebben wij een aanpak bestaande uit drie stappen toegepast voor het analyseren van mass-spectrometry-data om tot een discriminatieregel te komen (zie hoofdstuk 6). Peptide massa's werden gerangschikt met behulp van RF. Aangezien de variabelen die als hoogst gerangschikt werden bestonden uit sterk gecorreleerde variabelen, werden nieuwe variabelen aangemaakt om deze sterk gecorreleerde variabelen te groeperen. Deze nieuwe variabelen werden uiteindelijk geïncludeerd in een model om borstkankerpatiënten van controlepersonen te kunnen onderscheiden. Dit leverde een goed model op met een sensitiviteit en specificiteit van respectievelijk 86.8% en 85.7%.

Om iets aan het probleem van grote aantallen variabelen in nutrigenomic data te doen, bestaan er ook methodes die de hoeveelheid variabelen reduceren door middel van het creëren van een kleinere set van nieuwe variabelen. Bijvoorbeeld, partial least squares (PLS) reduceert het aantal variabelen door middel van het construeren van latente componenten. In hoofdstuk 7 hebben we PLS toegepast om de associatie van 83 plasma-eiwitten met coronaire hartziekte (CHZ) mortaliteit te bestuderen, alsook de associatie van deze eiwitten met intermediaire eindpunten die betrokken zijn bij de etiologie van CHZ (namelijk body mass index (BMI), HDL-cholesterol (HDL-C) en totaal-cholesterol. Op deze manier hebben we een set van 15 eiwitten geïdentificeerd welke 65% van de CHZ-mortaliteit kon voorspellen, alsook sets van eiwitten geassocieerd met BMI, HDL-C en totaal-cholesterol. Om een overzicht te krijgen van de relaties tussen geïdentificeerde eiwitten, intermediaire eindpunten en CHZ-mortaliteit hebben we vervolgens principale componenten analyse (PCA) als statistische methode toegepast. Deze toepassing liet zien dat eiwitten die betrokken zijn bij inflammatie de meeste variantie verklaarden, gevolgd door eiwitten betrokken bij het metabolisme en door eiwitten gerelateerd aan totaal-cholesterol.

In deze thesis zijn manieren van analyseren van grote nutrigenomic datasets behandeld hoe met de statistische moeilijkheden die in dergelijke data aanwezig zijn om te gaan. Met het toepassen van deze manieren op echte nutrigenomic datasets waren we in staat om duidelijk interpreteerbare statistische modellen alsook biologisch relevante informatie te verkrijgen. De resultaten hebben inzicht gegeven in de genen en eiwitten die betrokken zijn bij intermediaire eindpunten en bij complexe ziektes zoals borstkanker en CHZ. De volgende stap in nutrigenomic onderzoek is het combineren van data van verschillende biologische niveaus om meer complete biologische hypotheses te genereren betreffende de ontwikkeling van complexe ziektes bij mensen. Hopelijk zal nutrigenomics onderzoek in de nabije toekomst ook leiden tot de detectie van meer biomarkers met klinische relevantie. De vertaling van genomics onderzoek naar de volksgezondheid is tot nu toe beperkt geweest, maar nieuwe ontwikkelingen in het genomics onderzoek zijn veelbelovend en kunnen leiden tot toepassingen die de volksgezondheid verbeteren.

Abbreviations

Abbreviations

ACE	Angiotensin I converting enzyme
ADD	Adducin
ADRB	Adrenergic beta receptor
AGT	Angiotensinogen
AGTR	Angiotensin receptor
Apo	Apolipoprotein
BDNF	Brain derived neurotrophic factor
BMI	Body mass index
C	Complement
CA	Cancer antigen
CBS	Cystathionine beta synthase
CCR	Chemokine receptor
CEA	Carcinoembryonic antigen
CETP	Cholesteryl ester transfer protein
CHD	Coronary heart disease
CK-MB	Creatine kinase muscle brain
CNTF	Ciliary neurotrophic factor
CRP	C Reactive protein
CSF	Colony stimulating factor
CTLA	Cytotoxic T-lymphocyte-associated protein
EDTA	Ethylenediaminetertraccetic acid
EGF	Epidermal growth factor
ENA	Epithelial-derived neutrophil-activating peptide
F	Factor
FABP	Fatty acid binding protein
FCER1B	Fc fragment of IgE high affinity I receptor for beta polypeptide
FGB	Fibrinogen beta chain
GC	Group-specific component
G-CSF	Granulocyte colony-stimulating factor
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GNB	Guanine nucleotide binding protein (G protein) beta polypeptide
HDL-C	High-density lipoprotein cholesterol
ICAM	Intercellular adhesion molecule
Ig	Immunoglobulin
IGF	Insulin-like growth factor
IL	Interleukin
IL-1 ra	Interleukin-1 receptor antagonist
IL-5 ra	Interleukin-5 receptor alpha
ITGA	Integrin alpha
ITGB	Integrin beta
LDLR	Low density lipoprotein receptor
LIPC	Lipase, hepatic
LPA	Lipoprotein (a)

LPL	Lipoprotein lipase
LTA	Lymphotoxin alpha
LTC4S	Leukotriene C4 synthase
MCP	Monocyte chemotactic protein
MDC	Macrophage-derived chemokine
MIP	Macrophage inflammatory protein
MDR	Multifactor dimensionality reduction
MMP	Matrix metalloproteinase
MTHFR	Methylenetetrahydrofolate reductase
NOS	Nitric oxide synthase
NPPA	Natriuretic peptide precursor A
PAI	Plasminogen activator inhibitor
PAP	Prostatic acid phosphatase
PAPP-A	Pregnancy-associated plasma protein A
PON	Paraoxonase
PPAR	Peroxisome proliferator-activated receptor
PSA	Prostate-specific antigen
RF	Random forests
SAA	Set association approach
SCNN1A	Sodium channel nonvoltage-gated 1 alpha
SCYA	Small inducible cytokine A
SDF	Stromal cell-derived factor
SELE	Selectin E
SELP	Selectin P
SGOT	Soluble glutamic-oxaloacetic transaminase
SHBG	Sex hormone-binding globulin
TBG	Thyroxine-binding globulin
TCF	Transcription factor
TF	Tissue factor
TGFB	Transforming growth factor beta
TIMP	Metalloproteinase inhibitor
TNF	Tumor necrosis factor
TNF RII	Tumor necrosis factor receptor-like 2
Total-C	Total cholesterol
TSH	Thyroid-stimulating hormone
UCP	Uncoupling protein
UGB	Uteroglobin
VCAM	Vascular cell adhesion molecule
VDR	Vitamin D receptor
VEGF	Vascular endothelial growth factor
vWF	von Willebrand factor

Dankwoord – Acknowledgements

Dankwoord – Acknowledgements

Dit project was een droom met vele mooie momenten. Het was een plezier om deze droom te kunnen delen met andere mensen. Daarom wil ik iedereen bedanken die hier aan heeft bijgedragen.

Allereerst wil ik graag het Centre for Human Nutrigenomics en de mensen die hierbij betrokken waren bedanken voor het financieren van dit project. Ook mijn begeleiders Edwin, Edith en Jolanda bedank ik voor het vertrouwen dat ze in me hebben gesteld om dit project tot een goed einde te volbrengen. Jolanda, de eerste twee jaren op het RIVM waren een leerzame tijd, waarin je mij inwijdde in de ins en outs van het aio-schap. De dagelijkse begeleiding en de wekelijkse besprekingen heb ik veel aan gehad. Voor mij was het een plezier om met je om te gaan en met je samen te kunnen werken, je optimistische en vrolijke aard, maar ook de goede kwaliteit van werk die je nastreefde waarbij je grondig induiken van mijn werk tot vele verbeteringen van de manuscripten hebben geleid. Dank hiervoor! Edwin, de laatste twee jaren van dit project heb je mij begeleid. Ook al lag de nadruk soms meer op de statistiek dan de biologie, je hebt altijd meegedacht en waardevolle input gegeven. Je zachtmoedigheid, inzicht in omgang met mensen en humor zijn eigenschappen die ik waardeer. Deze eigenschappen bracht je ook naar voren in de wekelijkse bespreking met de Fungen groep, waarin iedereen zijn zegje kon doen en er naar iedereen geluisterd werd. Dit vond ik altijd een leuk moment van de week en gaf me ook het gevoel van betrokkenheid en onderdeel van deze groep te zijn. Dank je! Edith, het was een plezier om met je om te gaan, je enthousiasme voor het onderzoek en in de omgang met mensen maakten de besprekingen tot een plezier. Maar je had ook aandacht voor persoonlijke belevenissen, die de gesprekken aangenaam maakten. Veel dank voor het regelwerk voor de promotie, waaronder het samenstellen van de promotiecommissie, alsook dank voor de mogelijkheid om te kunnen blijven werken in Wageningen. Veel succes met je professorschap! Pieter van ´t Veer, dank je voor je bereidheid om als mijn promotor op te treden. Ook al bleek dit later niet meer nodig te zijn, je tips hebben o.a. geleid tot het volgen van een goede statistische cursus in Hasselt. Ook heb ik de samenwerking met jou en Jan Burema voor het vak exposure assessment als zeer prettig ervaren, dank jullie wel.

Mijn dank gaat ook uit naar de co-auteurs voor hun bijdrage. Graag bedank ik Nico Nagelkerke, Hans van Houwelingen en Uwe Thissen voor hun begeleiding in de statistiek. Nico, je bijdrage op afstand heeft geleid tot twee succesvolle publicaties. Ik vond het erg leuk dat je me indertijd hebt uitgenodigd om samen deel te nemen aan de classificatie competitie. Hartelijk dank voor de samenwerking. Uwe, de ontmoetingen op TNO en de kantine van de WUR hebben geleid tot een mooi stuk. De gemoedelijke omgang maakte de samenwerking ontspannen. Naast het uitvoeren van de PLS analyses heb je me uitleg gegeven in PLS en SVM, waardoor ik meer inzicht heb gekregen in deze technieken, dank je! Wendy, het overleggen, dingen uitzoeken, taktische antwoorden bedenken voor de reviewers etc. maakten de lange dagen op het Rikilt tot een leuke tijd. Het was een plezier om met je samen te werken! Jaap, naast het initiëren van de analyses van de microarray studie heb je ook aan het artikel zelf een grote bijdrage gehad. Je enthousiasme, ideëen en inhoudelijke kennis die je naar voren bracht tijdens het overleggen waren een goede stimulans. Ping, thank you for your kind help, I appreciate this. Wish you and Kaimei the

best! Daphne, dank je voor je goede begeleiding en de uitleg die je hebt gegeven. Ik vond het erg leuk om bij je binnen te kunnen lopen om iets te vragen of een persoonlijk praatje te maken.

De eerste 2 jaren van mijn project heb ik een fijne en leuke tijd gehad bij de afdeling CVG van het RIVM. Deze leuke tijd is mede te danken aan de collega's: Marion, Daphne, Saskia, Brian, Frederike, Janneke, Du, Joop, Hans, Martinet, Heidi, Jan, Elly, Jeljer, Maryse, Martine bedankt voor de gezellige lunches! Marion, je was een fijne kamergenote, waarbij je me goed aanvoelde. Ondanks de hitte van de kamer konden we lachen en was het gezellig. Verder dank ik ook de secretaresses voor hun hulp, in het bijzonder Karin van Mourik, dank je voor je warme belangstelling. Ook was ik blij met de hulp van Hans van der Westelaken bij computerproblemen en de hulp van Gerda Doornbos met SAS.

In mijn tweede jaar kreeg ik ook een werkplek in Wageningen bij de groep van Michael Müller, waar ik een paar dagen per week kon werken. Het was fijn dat ik deze mogelijkheid kreeg, ik had het naar m'n zin op deze afdeling. Saskia, het was leuk om af en toe een praatje met je te kunnen maken. Linda, it was a nice surprise to meet you again at Wageningen. Verder bedank ik Mark Boekschoten voor de hulp bij het vinden van een geschikt pathway programma voor m'n analyses. Anand, thank you for your help with accessing the computer cluster of Wageningen. It was a pleasure to talk with you. Kevin, it is nice to work together with you, you are a friendly and excellent PhD-student. I also thank the other PhD-students of Wageningen University. Du, thank you for your help and your friendship. The dinner in the chinatown of Boston was nice. Antonie, leuk om je te hebben leren kennen. Een mooie nieuwe tijd toegewenst met wat je gaat doen. Akke, bedankt voor je gezelligheid. Gerda, leuk dat we nog even kamergenoten kunnen zijn. Anastasia, thank you for the nice conversations, it is great to talk so open with you. Also thank you Rina for being my buddy during the PhD-tour, it was a great pleasure to meet you. Many, many thanks for the organizers of the PhD-tour in the United States and all the PhD-students who participated. It was a great journey together, with many beautiful moments. Thank you all!

Ook de laatste twee jaren bij de afdeling van Humane Biologie (HB) heb ik een leuke tijd gehad. Dit is te danken aan de collega's en in het bijzonder de Fungen group. Freek, je humor en ontspannen houding maakten de omgang aangenaam. Dank ook voor al je hulp, van analyst tot installeren van software en beantwoorden van computervragen. Janneke, het was leuk om af en toe bij je binnen te kunnen lopen en een praatje te maken over van alles en nog wat. Kaatje, je openheid tijdens gesprekken hebben me in positieve zin geraakt. Ook bedank ik natuurlijk Egbert, Ronny, Johan (Renes), Jonathan, Anja en Anke voor de leuke tijd samen bij HB! Verder bedank ik Loek en Paul voor de hulp met de computer als er iets aan de hand was, alsook de secretaresses van HB voor hun hulp bij praktische zaken. Ralph, bedankt voor de gezelligheid en je tips. Jos en Johan (de Vogel), leuk dat jullie er ook bij het afscheidsetentje bij waren. De meeste tijd bij HB heb ik doorgebracht met m'n kamergenoten: Sander, Sandy, Uriël en Floortje, bedankt voor de vele leuke momenten! Sander, je was een aangename kamergenoot met een positieve en goedlachse instelling. Je stond altijd klaar om mij en anderen te helpen bij vragen en als er iets te regelen viel voor anderen, dan was dat goed verzorgd. Veel dank en ik wens je succes met het afronden van je thesis en met alles wat je daarna gaat doen. Ook wens ik je een goede tijd toe samen met Patricia en Coen. Sandy, bedankt voor de vele goede zorgen die je aan de dag legde voor ons en de plant. Ook de geweldige nieuwtjes waarmee je ons iedere week opnieuw wist te

verrassen zal ik missen. Zullen we nog één kopje thee doen? Het etentje bij Guio's met Sander en Boris was een bekroning op de leuke tijd samen. Uriëll, je humor en verhalen zorgden altijd voor sfeer op de kamer. Ik hoor dat je een goede tijd hebt in Amsterdam, maak er wat moois van. Floortje, wat moet ik nu doen met mijn toilet…? Ik vond het altijd leuk om met je te kletsen. Niet te veel meer nadenken over die Bland-Altman plots, een goede nieuwe tijd toegewenst! Mandy, het was moeilijk om te weten wanneer je mij plaagde, maar ik weet nu dat dat eigenlijk altijd het geval was. Ik heb erg met je kunnen lachen, dank hiervoor. Hopelijk kun je er bij de promotie bij zijn! Boris, op mijn laatste dag ben je nog gepromoveerd tot kamergenoot, maar eigenlijk was je dat al de hele tijd. Ik heb genoten van de vele shockerende opmerkingen, gezellige koffie/theemomenten en de mooie schaakpartijen! Vera, Judith, Maartje en Marjet, jullie ook bedankt voor de gezellige koffiepauzes. Marcel, het was leuk om samen te werken op de SNP-analyses, maar ook buiten de analyses om was het gezellig, dank je.

Dear Young Nugies, thank you all for the unforgettable time we spend together in the beautiful places of Marseille, Tuscany, Oxford, and Oslo. Hope you are all doing well and wish you the best! Fre, many thanks for making this all possible, for your enthusiasm in guiding us, it was great to listen to the stories you could tell during conversations. Also, Frans Bouman, your input was a great stimulation to connect with each other.

Ed, Bart, Claudia, Rick, het kunnen lachen om elkaars gekkigheden, de avonden bij Bart's Boulderzolder, de uitgebreide franse ontbijten en natuurlijk niet te vergeten het gezamenlijk klimmen en boulderen waren om te genieten. Rick, ook veel dank voor het verzorgen van de cover. Cecile, het was fijn om samen te kunnen klimmen tijdens de avonden in klimhal centraal. Ook de klimmers en medewerkers (Lilly, Fonsa) van Rocca in Gulpen bedankt voor de leuke tijd! Naast het klimmen was het heerlijk om onderling grappen te maken. In het bijzonder dank ik Maarten, Joost, Erik, Martin en Bas voor de goede en leuke klimsessies.

Veel dank ook aan de groep van djembe-les, Marja, Huub, Jurgen, Mathilda, Anoek, met veel plezier heb ik met jullie samengespeeld. Dank ook aan Zulema en Peter voor het lesgeven, ik heb er erg van genoten! Marja, bedankt voor de aangename maandagavonden, het was leuk om samen djembe te spelen. Ook vond ik het erg leuk om samen te praten en te filosoferen.

Lex, dank je voor je vriendschap. Ik mag je graag en het is aangenaam met je te praten en samen dingen te ondernemen. Wanneer is de volgende wandeling? Maya, je komst naar Limburg en het samenzijn met z'n allen in Didam waren fijne momenten, dank je wel. Inge, leuk je te hebben leren kennen. Monica, dank je voor het stimuleren om te beginnen aan het aio-schap. Zou leuk zijn om je weer te zien! Krispijn, Rens, Wim, was leuk om elkaar weer te zien in Limburg. Ik hoop dat jullie naar de verdediging komen, maar dan wel met snor! Sezgin, je hebt me vaak gestimuleerd om dingen te ondernemen (student assistent, oppas, De Jaren, beginnen van aio-schap) die me hebben geholpen, dank hiervoor!

Thank you all 'Dutch' practitioners of Tensegrity for the moments of practicing together and sharing many moments of joy and beauty. I also thank the practitioners for dreaming together in organizing events: Ine, Ed, Carine, Arthur, Liesbeth, Colline, Marleen, Javor, Alga (gracias!), Shanti, Hildert, Vivhar, Ilonka, Jessica, Paul, Corry, Herman, Paula, Femke thank you all from my heart. Hildert en Vivhar, het samenwonen in Leiderdorp was een aangename tijd van leren en samen zijn.

In het bijzonder bedank ik mijn paranimfen Ed en Hildert. Ed, dank je voor je vriendschap, je humor en enthousiasme en de vele momenten van support. Het is leuk dat we zoveel gemeenschappelijke interesses kunnen delen. Het samen zijn en uitwisselen van ideëen en visies met jou is altijd stimulerend. Hildert, ik dank je voor de vele mooie en vreugdevolle momenten die we samen hebben gedeeld, bijvoorbeeld Barcelona, Karma Ling, de wandelingen in Utrecht en Limburg, alsook het samen koken en eten. Ik ben blij dat ik me zo op m'n gemak kan voelen als we samen zijn.

Verder bedank ik mijn buren Ruud en Carla in Cadier en Keer voor hun hulp en voor het in het oog houden van mijn huis op de tijden dat ik er niet was.

Simon, Raimond, Martijn, Robin, Noora, de mooie tijd die we samen hebben doorgebracht zal ik altijd in mij meedragen. Ik ben blij dat we elkaar onlangs weer hebben ontmoet, om te horen hoe het met jullie gaat en waar jullie je mee bezig houden.

Als laatst dank ik mijn familie, de Spreij-en en de Heidema's. In het bijzonder wil ik Adriaan bedanken, voor de leuke en aangename gesprekken, alsook voor het advies om te kiezen voor wat ik echt wilde, dit heeft me geholpen om voor dit aio-project te gaan. Clary en Hendrik, veel dank voor de vele leuke momenten die ik heb kunnen doorbrengen samen met jullie en Klaas-Hendrik op de boerderij. Klaas-Hendrik, ik kan het goed met je vinden en ook in het praten herkennen we elkaar. Lieve opa en oma Spreij, bedankt voor jullie oprechte liefde en de plezierige tijd die jullie me hebben gegeven, bijvoorbeeld samen wandelen in het bos, poffertjes eten bij de panoramahoeve en oneindig veel meer. Ook ben ik heel dankbaar dat jullie me altijd hebben gestimuleerd in het leren. Lieve Nanny, ik mag je graag en ben dankbaar om een lieve zus te hebben zoals jij. Ik denk regelmatig aan je en voel dat je altijd dicht bij me bent. Berry, het is leuk dat je bij de familie bent gekomen, je bent een fijne en aangename kerel.

Lieve papa en mama, jullie hebben me de mogelijkheid gegeven om tot dit punt in mijn leven te komen. Jullie onvoorwaardelijke steun en liefde raakt me diep in mijn hart. Ook de openheid die jullie hebben voor de dingen waar ik me mee bezig houdt ben ik heel blij mee. Ik houd heel veel van jullie en ben blij dat we zoveel mooie momenten met elkaar hebben mogen delen, bijvoorbeeld de mooie vakanties, maar ook de eenvoudige dagelijkse dingen zoals samen eten. Oneindig veel dank hiervoor. Ik hoop dat de toekomst nog vele mooie momenten zal brengen. Met veel liefde denk ik aan jullie.

Geert

About the author

Curriculum Vitae

Andries Geert Heidema was born on the 24th of March, 1977 in Ede, The Netherlands. In 1996 he started to study psychology at the Free University in Amsterdam. During his specialization in biopsychology he studied stress physiology, brain electrophysiology, and behavioral genetics. For his master thesis he performed research in the United States at the University of Vermont in Burlington. During this period he analyzed twin data to investigate the genetic and environmental contribution to behavioral disorders in adolescents. After his return he graduated in 2002 at the Free University in Amsterdam.

As the relation between food and well being is one of his interests, he worked between 2002 and 2004 in Amsterdam at an organic branch, learning more about food and diets. In 2004 he started this PhD-project, which has been funded by the Centre for Human Nutrigenomics. In the first two years of this project he worked at the Centre for Nutrition and Health at the National Institute for Public Health and the Environment (RIVM) and at the Division of Human Nutrition at Wageningen University. The last two years he carried out his PhD-project at the Department of Human Biology of Maastricht University.

List of publications

Articles

First author

- AG Heidema and N Nagelkerke. Developing a discrimination rule between breast cancer patients and controls using proteomics mass spectrometric data: a three-step approach. Statistical Applications in Genetics and Molecular Biology 2008, Vol. 7(2):5.
- AG Heidema*, W Rodenburg*, JMA Boer, IMJ Bovee-Oudenhoven, EJM Feskens, ECM Mariman, J Keijer. A framework to identify physiological responses in microarray-based gene expression studies: selection and interpretation of biologically relevant genes. Physiological Genomics 2008, 33(1):78-90.
- AG Heidema, EJM Feskens, PAFM Doevendans, HJT Ruven, HC van Houwelingen, ECM Mariman, JMA Boer. Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs. Genetic Epidemiology 2007, 31(8):910–921.
- AG Heidema, JMA Boer, N Nagelkerke, ECM Mariman, DL van der A, EJM Feskens. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genetics 2006, 7:23.
- AG Heidema, P Wang, CM van Rossum, EJM Feskens, JMA Boer, F Bouwman, P van ‘t Veer, ECM Mariman. Sex-specific leptin-independent effects of CNTF, IL6 and UCP2 polymorphisms on weight gain. Submitted.
- AG Heidema, U Thissen, JMA Boer, F Bouwman, EJM Feskens, ECM Mariman. The association of 83 plasma proteins with CHD mortality, BMI, HDL- and total-cholesterol in men: applying multivariate statistics to identify proteins with prognostic value and biological relevance. Submitted.

Other articles

- M den Hoed, AJPG Smeets, MAB Veldhorst, AG Nieuwenhuizen, FG Bouwman, AG Heidema, ECM Mariman, MS Westerterp-Plantenga, KR Westerterp. SNP analyses of postprandial responses in (an)orexigenic hormones and feelings of hunger reveal long-term physiological adaptations to facilitate homeostasis. International Journal of Obesity, in press.
- Y Lu, AG Heidema, P de Groot, MET Dollé, S Imholz, WMM Verschuren, C Wijmenga, M Müller, EJM Feskens, JMA Boer. Identifying additional genetic determinants, possibly with epistatic effects on plasma HDL cholesterol levels using random forests and interaction analysis. Submitted.

Presentation

- AG Heidema. A survey of multi-locus methods for analysis of large SNP datasets. Genetic Epidemiology II: Latest Developments. Henry Stewart Talks, 2007. Available at www.hstalks.com.

Abstracts

- AG Heidema, EJM Feskens, HC van Houwelingen, PAFM Doevendans, ECM Mariman, JMA Boer. Application of three approaches to select relevant SNPs in genetic association studies. IGES platform presentation, 2006.
- AG Heidema, EJM Feskens, HC van Houwelingen, PAFM Doevendans, ECM Mariman, JMA Boer. Application of three approaches to select relevant SNPs in genetic association studies. NUGO, 2006.
- AG Heidema, JMA Boer, DL van der A, ECM Mariman, N Nagelkerke, EJM Feskens. Overview of multi-locus methods for analyzing multiple SNPs. NUGO, 2005.

Training and supervision plan

Discipline specific activities

Courses
- Bioinformation Technology - 1, VLAG, Wageningen, 2004
- Modern statistical methods, Nihes, Rotterdam, 2004
- Gene expression meets genetic epidemiology, NUTRIM/VLAG, Maastricht, 2004
- NuGO Introduction course, NuGO, Marseille, 2005
- Advances in population-based studies of complex genetic disorders, Nihes, Rotterdam, 2005
- Masterclass Nutrigenomics, NuGO, 2005, 2006
- Data analysis and statistical techniques for SNP analysis, NuGO, 2006
- Analysis of longitudinal data, UM, Maastricht, 2007
- Multivariate data analysis, Diepenbeek, Hasselt, 2007

Meetings
- SNPs in diet-related disease, CHN, RIVM, 2004
- NuGO week, NuGO, 2004-2007
- Association analysis of candidate genes: from single point to multi-point methods, LUMC, Leiden, 2005
- Gene-nutrient interactions in the metabolic syndrome: a population-based epidemiological approach, NuGO, Amsterdam, 2005
- CHN PhD-project meetings, CHN, Wageningen, 2005-2007
- Multivariate statistical approaches for the analysis of nutrigenomic data, NuGO, Norwich, 2006
- IGES congres, IGES, 2006-2007
- The bioinformatics and statistical challenges associated with studies of multiple SNPs in nutritional studies, NuGO, Krakow, 2007
- Classification contest on clinical mass spectrometry based proteomic diagnosis, LUMC, Leiden, 2007
- Symposium “Chronic inflammation and obesity: friendly or hostile fire”, Div of Human Nutrition, CHN, VLAG, NGC, TIFN, Wageningen, 2007

General courses

- VLAG PhD week, VLAG, Bilthoven, 2005
- Philosophy and ethics of food science and technology, VLAG, Wageningen, 2006

Optional courses and activities

- Lectures at the RIVM, RIVM, 2004-2006
- Lectures at Wageningen university, WUR, 2004-2006
- Lectures at Maastricht university, UM, 2006-2008
- PhD study tour United States, WUR, Div of Human Nutrition, 2007

The research presented in this thesis was supported by the Centre for Human Nutrigenomics (CHN), The Netherlands.

The studies presented in this thesis were performed at the National Institute of Public Health and the Environment (RIVM) and within the Nutrition and Toxicology Research Institute Maastricht (NUTRIM). Both participate in the Graduate School VLAG (Food Technology, Agrobiotechnology, Nutrition and Health Sciences), accredited by the Royal Netherlands Academy of Arts and Sciences.

Cover design: Rick Verhoog, Parkers
Cover photography: Willem Doelman
Picture: Tensegrity structure by Buckminster Fuller
Printed by: Ponsen en Looijen, Wageningen, The Netherlands