Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 7

A Multiple Testing Approach to High-Dimensional Association Studies with an Application to the Detection of Associations between Risk Factors of Heart Disease and Genetic Polymorphisms

José A. Ferreira* Johannes Berkhof†

Olga Souverein‡ Koos Zwinderman**

*Vrije University Medical Center, Jose.Ferreira@rivm.nl

†Vrije University Medical Center, h.berkhof@vumc.nl

‡Wageningen University, olga.souverein@wur.nl

**University of Amsterdam, a.h.zwinderman@amc.uva.nl

A Multiple Testing Approach to High-Dimensional Association Studies with an Application to the Detection of Associations between Risk Factors of Heart Disease and Genetic Polymorphisms*

José A. Ferreira, Johannes Berkhof, Olga Souverein, and Koos Zwinderman

Abstract

We present an approach to association studies involving a dozen or so 'response' variables and a few hundred 'explanatory' variables which emphasizes transparency, simplicity, and protection against spurious results. The methods proposed are largely non-parametric, and they are systematically rounded-off by the Benjamini-Hochberg method of multiple testing. An application to the detection of associations between risk factors of heart disease and genetic polymorphisms using the REGRESS dataset provides ample illustration of our approach. Special attention is paid to book-keeping and information-management aspects of data analysis, which allow the creation of an informative and reasonably digestible 'map of relationships'—the end-product of an association study as far as statistics is concerned.

KEYWORDS: multiple testing, association tests, non-parametric methods

*We thank an anonymous reviewer for comments which led to the discussion in Section 2.2.

1 Introduction

One of the most common statistical questions arising in epidemiologic research is whether two or more characteristics or 'variables' that can be observed in individuals of a given population are associated. By saying that two or more variables are associated we mean that their 'underlying' joint distribution does not factorize into the product of their marginal distributions, and, accordingly, producing evidence of an association requires sampling from the population in question and assessing the relationship between joint and marginal empirical distributions in terms of a statistical test. To give an example, if one of the variables is HDL (high-density lipoprotein) cholesterol level and the other is the variant of a single nucleotide polymorphism, so that an individual has one of three genotypes, then a statistical test of association may be based on the comparison of the three distributions of HDL in the population strata of genotypes 1, 2 or 3, and the evidence against the hypothesis of no association may be weighed by means of the p-value—the probability that a sample drawn from the same population would, were there no association, yield a discrepancy between the three distributions bigger than or as big as the discrepancy actually observed.

The aim of epidemiologic research is often to establish causal relationships between variables as a means of identifying the main causes of a disease, and this is a long way from the mere detection of associations. For instance, even if an association between HDL and a polymorphism is established beyond reasonable doubt, it may very well be that the polymorphism is associated with another polymorphism that regulates the synthesis of a protein which in turn influences HDL, and it may be the identification of the latter association that turns out to be useful in developing new treatments for heart disease. Thus, as Rothman (2002) insists, epidemiology is "more than the application of statistical methods to the problems of disease occurrence and causation". On the other hand, the search for associations is usually a crucial first step in an epidemiologic investigation. This is especially the case in the study of so-called 'complex diseases' such as diabetes, heart disease, many forms of cancer, and neurological diseases—diseases attributable to a combination of several genetic and environmental factors—where, understandably, the starting point of the investigation often amounts to a collection of more or less speculative hypotheses about a large number of phenotypic and genotypic variables and further progress rests partly on a reliable picture of the statistical associations between those variables. Once good evidence is obtained about a number of associations, epidemiologic research will, ideally, proceed with the sketch of a provisional biomedical theory explaining the

causes of the disease and the formulation of hypotheses that can be used to test it, and eventually modify it in later stages, with new data (see p. 697 of Willett (2002) for specific examples). Often, as pointed out on p. 700 of Rees (2002), medical treatment is suggested by associations between variables long before any mechanistic explanation of the disease is advanced, and it is the effectiveness of the treatment that subsequently helps building a satisfactory theory. In any case, however long and tortuous the research path may be starting from the investigation of the first associations, and irrespective of whether it will eventually lead somewhere, it is clear that wrong tracks indicated by false positive results should be avoided; for if it is generally difficult to build a biomedical theory it is certainly more difficult to build a valid one out of the wrong elements. This is why the detection of *genuine* associations by means of statistical methods, despite being only a tiny, first step towards the biomedical elucidation of a disease, plays such an important role in the epidemiologic research of complex diseases (see also p. 697 of Willett (2002)).

Unfortunately, detecting associations between variables is not a completely straightforward task; the existence of a plethora of contradicting, irreproducible or downright spurious associations in biomedical science, now generally acknowledged and partly explained by some (Winkelmann et al. (2000), Willett (2002), Ioannidis (2005), Goodman and Greenland (2007)), is perhaps the best illustration of that. Besides scientific malpractices, there are at least three properly methodological or statistical (despite the fact that in most cases associations are tested with standard statistical tests) difficulties inherent to association studies that account for the abundance of false positive results. First, there is the problem of confounding, on which we will expand below; this is often responsible for distorting, in both magnitude and direction, associations between variables, and can even yield spurious results.

Secondly, there is the problem of multiple testing: in studies involving many tests, the practice of discrediting any null hypothesis whose p-value is below 0.05 is bound to give rise to many false positives within a single study.[1] Although such practice does not really deserve the qualification of 'malpractice' (in contrast to the repeated, unreported testing of hypotheses until 'significant' results are obtained, one of the forms of what may be called 'data dredging') it is becoming difficult to tolerate following the impact of multiple testing methods, such as Benjamini and Hochberg's (1995).

Finally, there is the problem of 'incorrect model specification'. In many association studies, hypotheses of association are tested within parametric

[1]The best part of Ioannidis's (2005) argument may be translated into the proposition that rejecting hypotheses on the basis of a fixed 'p-value threshold' necessarily yields many false positive results.

DOI: 10.2202/1544-6115.1420

regression models which specify a relationship between a 'response variable' such as HDL and several 'explanatory variables' such as the genotype of a polymorphism, age, body mass index (BMI), etc. (e.g. Kathiresan et al. (2007)). The rationale for doing this is that tests based on parametric models are more powerful than non-parametric tests and that by including other explanatory variables one simultaneously 'corrects for confounders' that may distort the association between the response and the explanatory variable of interest and (thanks to the decrease in noise achieved by the incorporation of greater 'explanation') further increases power. Despite the lucid appraisal of regression models made by some epidemiologists (e.g. Rothman (1986) in his Chapter 14), this point of view has become so widespread in biomedical research that it may rightly be called a paradigm.[2] And yet, in many cases—especially in cases where there are more than just a couple of variables and only a few hundreds of patients—the assumption that a regression model describes the relationship between response and explanatory variables sufficiently well so as to allow a reliable investigation of the association between the response and a particular explanatory variable is problematic.

In this work we propose an approach to association studies involving relatively large numbers of variables that obviates, or at least mitigates, the problems now described, and illustrate its application to the detection of associations between risk measures of heart disease and genetic polymorphisms. Our starting point is the assumption that the first aim of an association study—or of a 'low-level association study'—is to *detect* associations between a 'small' set of response variables and a 'large' set of explanatory variables—not to *quantify* or model putative associations, which we regard as a much more difficult, 'higher level' problem to be tackled in later stages when more concrete hypotheses have been formulated and more refined studies (e.g. based on more specific population groups) can be set up. The amount of research dedicated to establishing associations between two or more variables and the theoretical framework provided by some epidemiologists (e.g. Willett (2002)) justify this basic assumption. The response variables are to be thought of as risk measures for a certain disease—as for instance HDL cholesterol and blood pressure are for heart disease—and the explanatory variables as a set of genotypic and/or phenotypic variables—the genotypes of polymor-

[2]According to the experience of the first author, the situation has gone so far that during the process of reviewing papers one is sometimes asked to justify the use of a simple analysis based on non-parametric methods—for instance the application of the Wilcoxon-Mann-Whitney test within population strata—in place of a multivariate analysis based on a complicated parametric model, when it is the latter rather than the former that requires justification.

phisms, age, BMI, etc.—that could influence the response variables and thus possibly explain the occurrence of the disease.

Despite being based largely on standard statistical methods, the methodology we suggest contains original elements and departs from the run-of-the-mill association study in essential points. On the other hand, like most other approaches to association studies, it is fully based on p-values. Thus, given the philosophical objections sometimes raised to the use of p-values, it seems well to observe that the detection of associations by means of multiple testing methods typically achieves a very concrete purpose: in the case of the Benjamini-Hochberg method advocated here, it imposes an upper bound on the so-called *false discovery rate* and (roughly speaking) still finds as many genuine associations as is feasible.

Finally, it is important to state one thing that our approach is not: it is not an approach to prediction, nor to variable selection, nor to both. By 'prediction' we mean the activity of *guessing* the value of a response variable (HDL levels in a patient, say) on the basis of a set of explanatory variables (age, BMI, 'genetic profile', etc.), by means of a function—the predictor—relating the latter to the former; by 'variable selection' we mean the activity of selecting the explanatory variables that, used in the predictor, and according to a criterion such as 'mean-squared error' or 'probability of correct classification', best predict the value of the response variable. Trivially, an explanatory variable may be associated with a response variable but be practically useless in guessing its value; and an explanatory variable may be very accurate in guessing the value of a response variable without there being any *relevant* causal relationship between the two. Irrespective of whether a set of explanatory variables is useful to predict a response variable or is 'selected' on the basis of a criterion, establishing an association between the two can be very useful and remains a legitimate aim. Thus, as pointed out on p. 696 of Willett (2002), the information carried by a single phenotypic variable 'measured in the clinic' often subsumes the information provided by several polymorphisms, which renders the latter practically useless for prediction purposes; nevertheless, establishing that an association between the phenotypic variable and the polymorphisms exists can be crucial in establishing causality.

Section 2 presents and justifies our approach in more or less formal terms and, along the way, touches on the subjects of confounding, multiple testing and incorrect model specification; the Benjamini-Hochberg method, which is the basis of our approach, is described in Subsection 2.1, and Subsection 2.2 discusses possible pitfalls. The rest of the paper, Section 3, describes an application to the problem of finding associations between risk measures

DOI: 10.2202/1544-6115.1420

of heart disease and genetic polymorphisms using data from the REGRESS study (Jukema et al. (1995)).

2 Description and justification of the method

Let us recall the type of association study we have in mind: given a 'small' number of response variables of interest—cholesterol levels, blood pressure, etc.—and a 'large' number of explanatory variables—genetic polymorphisms, age, BMI, etc.—the objective is to look for associations between the response variables and the explanatory variables and to come up with a 'map of relationships' summarizing the main findings and which can be used to formulate a theory (no matter how rudimentary) explaining how the response variables are influenced by (and perhaps influence) some of the explanatory variables. In this setting, we propose an analysis in four steps. In the **first step** we test the association between each pair of response variables, rank the pairs according to their p-values (which reflect the strength of the associations), use the ranking to construct a graph—a dendogram obtained by a cluster analysis method, for example—summarizing the relationship between the response variables, and finally pick some of these variables to be used in the remainder of the analysis.

The purpose of this first step is *not* to establish causal relationships between the response variables; this would be virtually impossible using statistical methods—and certainly cluster analysis methods—alone. Its purpose is rather to give an overview of the response variables, of the putative relationships between them, and of the relative strength of the relationships that are most likely to exist, and eventually to help selecting a smaller set of response variables on which to focus. Typically, many of the response variables are *known* to be associated, and the *interesting* response variables have been fixed in advance. Thus, to exemplify, systolic and diastolic blood pressures are known to be strongly associated variables, as are age and cholesterol levels, and in a study aiming at detecting genotypes that contribute to heart disease it would be a matter of course to include risk measures of heart disease such as HDL and LDL (low density lipoprotein) cholesterol and blood pressure as response variables; however, if it were found that in a given group of patients HDL and 'triglyceride levels' had a very strong relationship, then one might consider, for simplicity's sake, discarding one of these two variables from the remainder of the analysis. Another purpose of the overview of the response variables and their relationships obtained in the first step is to help creating the 'map of relationships' in the final stage of the association study.

In the **second step** we try to detect associations between each one of the response variables (fixed in the first step) and the large number of explanatory variables by means of the Benjamini-Hochberg method, or, in those cases where this is deemed inapplicable, of the Bonferroni method; see Subsection 2.1 for a description of these methods. The main concern here is to avoid false positive results—by fixing an upper bound on the false discovery rate in the case of the Benjamini-Hochberg method, and on the probability of at least one false positive result in the case of the Bonferroni method. For this reason, it is important that the p-values be computed correctly and accurately. Parametric tests and tests based on asymptotic results yield numerically accurate p-values, but such p-values can be incorrect because, strictly speaking, they are computed under the incorrect null distribution; this is typically the case with t- and F-tests, and with tests on the value of a parameter in a multiple regression model (e.g. modelling HDL levels as a function of a polymorphism with age, BMI and sex as explanatory variables). While it is true that when carrying out a single test one can sometimes transform variables and/or choose a more appropriate model in order to get a reliable p-value, in a situation where many tests need to be performed the possibility that the p-values are biased presents a more serious problem, if only because the p-values that are of any consequence are usually very small and a violation of a model's assumptions typically expresses itself more seriously 'in the tails'. Of course, the amount of bias in a multiple testing procedure due to a bias in the p-values will depend very much on the type of data at hand, but as a general rule it seems sensible to consider using exact, non-parametric tests.

One disadvantage of exact, non-parametric tests is that the accurate calculation of p-values (with a number of decimal places that allows us to distinguish 0.001, say, from zero), which is usually based on simulation, may be time consuming; however, given the resources currently available to the average statistician, this does not appear to be a serious disadvantage.

Perhaps the argument more frequently invoked against the use of non-parametric tests is that they entail 'loss of power'—relative, one presumes, to a test based on a parametric model that fits the data sufficiently well. From a 'distributional' point of view—ignoring the potential benefits of correcting for confounder variables—statistical folklore based on the comparison between the t-test and the Wilcox-Mann-Whitney test and similar examples suggests that the purported loss of power is overrated and, in view of the more or less serious deviations from model assumptions that are likely to occur, often turns into a gain in power.

DOI: 10.2202/1544-6115.1420

Correcting for confounders can be advantageous in terms of power. This is obviously the case when a response variable is associated with an explanatory variable and the association is direct within one age stratum (say) and inverse within another, since neglecting age will then dilute the association in the overall population and consequently decrease power. On the other hand, it is equally obvious that there can be no theorem stating whether it is better to account or not account for confounders in general: if the association between the response and the explanatory variable varies in intensity but has the same direction across age strata, then we may well be better off, regarding power of detection, looking at the overall population. Thus it seems sensible to keep an open mind on the subject of correcting for confounders and to consider Rothman's (2002) recommendation of stratifying the population according to variables thought to distort, and especially invert, associations. Once strata have been defined, one can carry out independent multiple testing procedures within strata and compute a bound on the false discovery rate as the sum of the bounds on the false discovery rates within strata; or, alternatively, one can compute 'single p-values' as functions of the p-values from individual strata (e.g. using Fisher's method or by averaging the absolute values of test statistics across strata), and apply a multiple testing procedure to these.

Correcting for confounders *in a regression model* can also increase power, both relative to an approach that simply ignores confounders and to an approach that stratifies the population according to confounders—the 'nonparametric approach to correcting for confounders' recommended by Rothman (2002) which we have just mentioned. However, one can only hope to 'correct for confounders' with a model that is approximately correct, which brings us back to the objections to parametric models raised four paragraphs above: if one agrees that biased p-values should be avoided in a multiple testing problem and that it is generally difficult to assess the bias incurred in the calculation of many p-values under a parametric model, then one should at least consider *not* using tests based on regression models that correct for confounders. It is interesting to point out that in association studies the incorporation of a confounder such as age or BMI in a regression model often does not account for interactions between the explanatory variable of interest (e.g. a polymorphism) and the confounder (e.g. Kathiresan et al. (2007)), while one would expect the parameter corresponding to the former to depend on the values of the latter (for example, we shall see towards the end of our application in Section 3 that the effect of certain polymorphisms on HDL is likely to depend on age). This practice is understandable in view of the number of parameters one would have to have in a model including main ef-

fects of confounders and enough interactions with the explanatory variable of interest, but it raises the question of whether regression models can generally be expected to yield 'legitimate' p-values in a multiple testing context.

At this point it is important to state two advantages that our approach has over the more standard approaches to association studies. In the first place, the use of non-parametric methods, including the eventual correction for confounders by stratification, makes the results of the study rather robust to the particularities of the analyst: for instance, there are only a few 'right' non-parametric methods of testing an association, and those few ones are usually comparable regarding both power and assumptions. This robustness is obviously helpful in isolating the causes of eventual spurious associations and inconsistent results. In contrast, an approach based on regression models leaves so much room for manoeuvre that it is frequently difficult to compare the results of two independent studies on the same problem.

Secondly, the combination of non-parametric tests with a multiple testing procedure makes the results of the study essentially *unique*. In particular, the fact that all associations between a response variable and the explanatory variables are tested ensures that no 'non significant' results are left unreported and that the evidence in favour of purported associations has been properly weighed. In contrast, the validity of a study in which only a small subset of explanatory variables is considered, largely dispensing the need for a multiple testing method, rests on the assumption that no other explanatory variables were investigated. Of course, in some studies it makes perfect sense to study associations between a response variable and a small set of explanatory variables, but these are 'high level studies'; in the 'low level studies' we consider here, where there are many more explanatory variables than response variables, it is unlikely to have each response variable coupled to a small set of explanatory variables.

To go on with the description of our approach, at the end of the second step we will have in principle declared significant some of the associations between the response variables and the explanatory variables; in the **third step** we focus on the explanatory variables for which at least one association with a response has been singled out and try to detect pairwise associations between them, again using non-parametric tests and a multiple testing procedure. The associations tested for in the third step are thus conditional on what was declared significant in the second. One might well consider testing the pairwise associations between *all* explanatory variables, but this seems less interesting for two reasons: first, increasing the number of hypotheses typically decreases power of detection; secondly, concentrating on the explanatory variables which are likely to be related to the response variables

DOI: 10.2202/1544-6115.1420

will lead to a smaller and more connected 'map of relationships', which is perhaps more helpful as a starting point for building up explanations than a more inclusive but sparser one.

Finally, in the **fourth step** we summarize and *verify* the associations declared significant in the second or third steps. The simplest and most faithful summary of the results will be a graph where each node represents a response variable or a set of explanatory variables and an edge linking two nodes indicates an association that was declared significant. But in drawing such a 'map of relationships' one should be allowed a certain amount of freedom to add response variables other than those considered in the second and third steps and to cluster together explanatory variables that are likely to be associated with each other and associated with the same response variable, for example. There is hardly any danger of overinterpreting the findings by doing this, as edges and clusters are determined by associations that were declared significant.

Post hoc information concerning possible confounding or 'effect modifying' variables may also be added to the graph. Such information can be gained by *verifying* the associations, namely by examining the distribution of a response variable as a function of the relevant explanatory variable across strata of a third variable thought to influence the direction or intensity of the association between the response and explanatory variables. In order to avoid the obvious problems of so-called 'sub-group analyses' it is important to regard any *post hoc* observation as tentative and the whole process of 'verification' as exploratory—and not to attach too much value to '*post hoc* p-values'. The recognition that certain apparently correct observations cannot really be substantiated by means of a p-value does not really neutralize them; if those observations have been made and can be properly understood within a biomedical context then they will carry their weight into the next stage of research.

In order to motivate further this fourth step, and in particular the 'verification of associations', we need to elaborate a bit on the subject of confounding. Suppose that the result of a statistical test indicates that HDL levels and a genetic polymorphism are associated, the two variables being modelled as two random variables X and Y, and think of a collection of other variables, such as age, BMI, smoker status, etc., modelled jointly with X and Y as a random vector Z. For simplicity, assume that Y takes the values 1 and 2 and that X and Z are discrete. The result of the test indicates that

$$\begin{aligned} 0 &\neq \Delta := E[X|Y=1] - E[X|Y=2] \qquad (2.1) \\ &= \sum\nolimits_z E[X|Y=1, Z=z] f_{Z|Y}(z|1) - E[X|Y=2, Z=z] f_{Z|Y}(z|2), \end{aligned}$$

in the usual notation of conditional probability functions and expectations. If $f_{Z|Y}(z|1) = f_Z(z) = f_{Z|Y}(z|2)$ for all z then the result of the test indicates that

$$0 < |\Delta| \leq \sum\nolimits_z |E[X|Y = 1, Z = z] - E[X|Y = 2, Z = z]|\, f_Z(z),$$

hence that $|E[X|Y = 1, Z = z] - E[X|Y = 2, Z = z]| > 0$ for some z; in other words, if Z is independent of Y then an association between X and Y is *genuine*, at least in the face of Z, in the sense that a difference in the means of Y in the two groups determined by Y exists within some of the strata determined by Z. This observation suggests a method of checking for spurious associations: once evidence in favour of an association between X and Y has been found, one can examine—graphically, in an exploratory fashion—the distribution of X across strata determined jointly by Y and Z and check—by means of *post hoc* tests—whether the purported association is visible in at least some of those strata.

It may happen that the association between X and Y has roughly the same direction and intensity in all strata, in which case Z would not be confirmed as a confounder. More frequently, however, the association between X and Y will have the same direction but vary in intensity across the strata of Z; for instance, if Y is a genetic factor and Z is age, then the effect of Y on X is expected to weaken with increasing age (cf. p. S19 of Winkelmann et al. (2000)). In this case it may be concluded that $E[X|Y = 1, Z = z] - E[X|Y = 2, Z = z]$ is a function of z and one will regard Z not only as a confounder but more specifically as an 'effect modifier'—the effect subject to modification being the difference in means of X in the two groups determined by Y. Less commonly, the effect modification will take the form of an inversion, $E[X|Y = 1, Z = z] - E[X|Y = 2, Z = z]$ appearing to be positive for z within one stratum and negative within another ('Simpson's paradox' is a standard example illustrating *the consequences* of this situation). Such an observation will normally need to be strengthened by a plausible biomedical explanation.

Irrespective of whether $E[X|Y = 1, Z = z] - E[X|Y = 2, Z = z]$ appears to depend on z or not—irrespective of possible confounders—in any of the three events just described there are good grounds for declaring the association to be genuine. On the other hand, in the event that the association between X and Y 'evaporates' when viewed across the strata determined by Z—if it appears very weak and to vary randomly in direction, say—then it seems sensible to assume that $E[X|Y = 1, Z = z] = E[X|Y = 2, Z = z]$ for all z, and hence, by (2.1), that

DOI: 10.2202/1544-6115.1420

$$0 < |\Delta| \leq \sum\nolimits_z \left| f_{Z|Y}(z|1) - f_{Z|Y}(z|2) \right| E[\left| X \right| | Y = 1, Z = z];$$

but since this implies $f_{Z|Y}(z|1) \neq f_{Z|Y}(z|2)$ for some z, we see that a failure of an association between X and Y to materialize within strata of a putative confounder Z provides evidence of an association between Z and Y and hence that Z *is* a confounder.

Thus, to sum up, the *post hoc* verification of associations is both a means of strengthening the evidence in favour of an association and a means of explaining why an association is likely to be spurious, and any further evidence derived from it can in principle be incorporated in the 'map of relationships' (examples are given at the end of Section 3). Of course, in theory the vector Z might contain so many possible confounders as to render any *post hoc* analysis useless; but in practice one must simply be reconciled with the fact that an association can only be verified in the face of a few obvious confounders and that any causal interpretation of the 'map of relationships' is tentative.

2.1 The Benjamini-Hochberg method

The Benjamini-Hochberg method is designed to test many, say m, null hypotheses simultaneously, of which an unknown proportion γ is true. It consists of computing a set of m test statistics—one for each hypothesis—and the corresponding set of m p-values (which requires the knowledge of the null distribution in each case), and then rejecting those null hypotheses whose p-values fall below a *data-dependent threshold*. This threshold, given by $X_{R_m:m}$, where $R_m = \max\left\{i : X_{i:m} \leq q\frac{i}{m}\right\}$ and $X_{1:m}, \ldots, X_{m:m}$ denote the ordered p-values and q is a number between 0 and 1 chosen by the user, has been designed so as to keep the expected ratio of the number of incorrect rejections (or 'false positives') to the number of rejections, called **false discovery rate**, below $q \times \gamma$. Since $q \times \gamma \leq q$, and since γ is usually unknown in practice and can be arbitrarily close to 1, what one can generally assert is that the *Benjamini-Hochberg method controls the false discovery rate at* q; for this reason we will refer to q as the **conservative (bound on the) false discovery rate**. However, in some cases one is able to find an estimate of, or at least a plausible upper bound for, γ, call it $\bar{\gamma}$, and thus make the sharper statement that the Benjamini-Hochberg method controls the false discovery rate at $q \times \bar{\gamma}$. In this form, the method is usually referred to as the **adaptive Benjamini-Hochberg method**, though there are several variants of it, each based on an estimator of $\bar{\gamma}$ (see the subsection below).

The interpretation of q is somewhat analogous to that of the significance level in a standard testing procedure, but not quite. What needs to be borne in mind is that the bigger q, the bigger the number of rejections and the bigger the number of *correct* rejections (i.e., the number of false null hypotheses one successfully identifies); but the bigger q, the bigger the false discovery rate. The logic of the Benjamini-Hochberg method is to fix the fraction of false positives among the 'positive' results one comes up with at a tolerable level (dictated by the user), and hope that such level will allow the discovery of many true positive results.

The Benjamini-Hochberg method is more powerful than traditional multiple testing methods like the Bonferroni method. Recall that the **Bonferroni procedure** of level α consists of rejecting those null hypotheses whose p-values fall below the *deterministic threshold* α/m, which guarantees that the chance of unduly rejecting one or more hypotheses is at most α. Intuitively, if m is very large then α/m is very small, which will typically yield very few rejections (very few 'discoveries'). In contrast, the threshold provided by the Benjamini-Hochberg method typically approaches a *positive* number as m increases, and therefore yields a sizable set of rejections.

Despite its advantages and general applicability, the Benjamini-Hochberg also relies on certain assumptions. The control promised by the method is exact only under certain conditions on the dependence structure of the sample of p-values. If m is large and $\gamma > 0$ (which is the case in the type of applications we have in mind) then the method will work in the approximate sense that the **false discovery proportion**, the ratio of the number of incorrect rejections to the number of rejections (whose expected value is the false discovery rate), is bounded above by approximately $q\gamma$ with high probability—irrespective of whether the p-values are independent or not—provided *one condition* is fulfilled.

To see what this condition is and understand what 'approximate' means we need to introduce some notation. First, let $T_1, T_2, \ldots, T_m$ denote the test statistics and assume that the test statistics computed under the alternative hypotheses tend to take smaller values than those computed under the null hypotheses; in particular, p-values based on unbiased tests are test statistics with this property. In order to distinguish the test statistics computed under the null hypotheses from the rest assume that the first $m_0 = \gamma m \leq m$ statistics are computed under the null and denote them by $T'_1, T'_2, \ldots, T'_{m_0}$. Finally, assume that $T'_1, T'_2, \ldots, T'_{m_0}$ all have the same distribution function, which we denote by F.[3] Typically, the test statistics are the p-values and F is the uniform distribution function on $[0, 1]$, but in principle F can be anything and does not even need to be continuous.

DOI: 10.2202/1544-6115.1420

Now consider the multiple testing procedure that consists of rejecting all the test statistics strictly below a given threshold t_m (or, more precisely, of declaring false all the hypotheses whose statistics are $< t_m$); the Benjamini-Hochberg method is an example of such a procedure based on the random threshold $X_{R_m:m}$. By definition, the false discovery proportion incurred by this procedure is either zero (in case no rejections are made) or equal to

$$\frac{\text{no. of incorrect rejections}}{\text{total no. of rejections}} = \frac{m_0}{m}\frac{\frac{1}{m_0}\sum_{i=1}^{m_0} 1_{[T_i'<t_m]}}{\frac{1}{m}\sum_{i=1}^{m} 1_{[T_i<t_m]}} = \gamma\frac{F_m(t_m-)}{H_m(t_m-)},$$

where 1_A denotes as usual the indicator function of the event A (equal to 1 if A occurs, equal to 0 otherwise), H_m is the empirical distribution function of the test statistics, and F_m and is the empirical distribution function of the test statistics computed under the null (the first m_0). It follows from this expression that if t_m is defined by[4]

$$t_m = \sup\{t:\ F_m(t) \le qH_m(t)\} \tag{2.2}$$

then the false discovery proportion is bounded above by $q\gamma$. Consequently, the procedure based on the threshold defined by (2.2) controls the false discovery proportion at q. Moreover, this procedure is optimal among all procedures that consist of rejecting all test statistics below a given threshold and that make no assumptions about the value of γ, because (i) conditionally on the q chosen, the higher the threshold the bigger the proportion of correct rejections, and (ii) increasing the threshold above the t_m of (2.2) is impossible without violating the upper bound $q\gamma$ on the false discovery proportion.

Unfortunately, the optimal procedure is not realizable in practice because F_m, unlike H_m, is not observable. However, since each of $T_1', T_2', \ldots, T_{m_0}'$ has distribution function F and $EF_m(t) = F(t)$ for all t, there is reason to hope that F_m is relatively close to F and consequently that

$$t_m' = \sup\{t:\ F(t) \le qH_m(t)\} \tag{2.3}$$

is close to the t_m of (2.2). If that is the case, then the set of hypotheses declared false by the multiple testing procedure based on t_m' will closely coincide with the set of hypotheses declared false by the multiple testing procedure based on t_m, and the procedure based on t_m' will not only be close to controlling the false discovery proportion at q but will also be close to being

[3]In practice, the correct calculation of a p-value requires the knowledge of the null distribution of the test statistic, and with this knowledge one can apply a transformation to the test statistic in order that its distribution function be (approximately, if the statistic is discontinuous) equal to F; hence this last assumption can be met in most situations.

optimal among similar procedures. And, as it turns out, the procedure based on (2.3) is, in essense, the Benjamini-Hochberg method, for it can be shown that when F is the standard uniform distribution function t'_m coincides with the threshold $X_{R_m:m}$ defined earlier.

In conclusion, the condition required for the Benjamini-Hochberg method to work in an approximate sense is that *the empirical distribution function of the test statistics under the null be close to their common distribution function*. And since under very general conditions on the dependence structure of $T'_1, T'_2, \ldots, T'_{m_0}$ F_m converges uniformly to F in probability as $m_0 \to \infty$ (e.g. Ferreira and Zwinderman (2006), Section 3.1), it is correct to state that *when m is large the method works very generally, albeit in an approximate sense*. This statement has been amply illustrated by simulation studies based on a wide variety of dependence structures (e.g. Benjamini et al. (2006), Kim and Wiel (2008), Romano et al. (2008)). As with many asymptotic results in statistics, however, its generality is also the source of its weakness, for the adjectives 'close' and 'approximate' are difficult or impossible to quantify in actual practice. For this reason, besides invoking arguments of plausibility as to why the Benjamini-Hochberg method should work in each case one must also try to check or justify the approximate uniformity of the empirical distribution of the p-values computed under the null; as will be seen in Section 3, this can be done using simulation and the data themselves.

2.1.1 Adaptive versions

As explained before, if one can find an estimate or upper bound $\bar{\gamma}$ for the proportion of true null hypotheses then, everything else being in place, the Benjamini-Hochberg method controls the false discovery rate at $q \times \bar{\gamma}$, a more powerful statement than that of the 'standard' method. Several methods of estimating or majorizing γ have been proposed (Benjamini and Hochberg (2000), Storey and Tibshirani (2003), Storey et al. (2004), Langaas et al. (2005), Ferreira and Zwinderman (2006), and Benjamini et al. (2006)), but most of them can be regarded as estimates of the *left-hand limit of the density of the p-values at* 1. The 'overestimator' of γ we shall use here is the left-hand limit of the histogram (based on Scott's rule) of the p-values at 1. Under general conditions (e.g. Section 4 of Ferreira and Zwinderman (2006)), this type of overestimator approaches a constant that exceeds γ with high probability as m increases; consequently, the false discovery proportion is bounded above by $q \times \bar{\gamma}$ with high probability for large m, and so is its expectation, the false discovery rate.

[4] If no t satisfies the condition within braces, $t_m = -\infty$ and no hypotheses are rejected.

DOI: 10.2202/1544-6115.1420

Of course, in practice one can neither ascertain the validity of the asymptotic results nor the accuracy of the desired bound on the *actual* false discovery rate with the m at hand, and so, ultimately, the application of the Benjamini-Hochberg method requires a leap of faith (in contrast with Bonferroni's, for example, which is as reliable as the p-values it is based on). Simulation studies (e.g. Kim and Wiel (2008), Dudoit et al. (2008)) show that the estimators of γ and the actual false discovery rate incurred in the adaptive Benjamini-Hochberg method are generally conservative (as they are intended to be) under a wide range of data-generating models, and that, when biased upwards, they do not dramatically exceed their nominal values. Given the theoretical background and such observations, simple 'diagnostic' procedures (see the text around Figure 7 in Subsection 3.2) can help further to motivate and justify applications of the Benjamini-Hochberg method.

2.2 Discussion

One of the reviewers raised two interesting points about the global control of the false discovery rate. Since there are at least as many multiple testing problems as response variables, and since the ultimate aim is to create a 'map of relationships', it may be argued that the control of the false discovery rate should be imposed on the whole set of tests rather than separately for each multiple testing problem. Of course, it is undesirable to simply pool together tests pertaining to different variables (for instance, if one response variable has more associations with the explanatory variables than another, then a 'pooled' multiple testing procedure will tend to detect only associations pertaining to the first, and yet the second variable is supposed to be important in its own right), so it is not appropriate to apply the Benjamini-Hochberg method to the union of the sets of p-values. Instead, one can compute (as in the case of the multiple testing problems by strata mentioned in Section 2) a bound on the 'global' false discovery rate as the sum of the bounds on the individual false discovery rates. Unless there are only a couple of response variables, however, this procedure will tend to be conservative. Alternatively, if one is in a position to take the 'leap of faith' mentioned in the previous subsection, one can go for the sharpest approach: If in the multiple testing problem pertaining to the i-th response variable we reject R_i hypotheses at a conservative false discovery rate of 0.1, then it can be argued that among the R_i rejected there are roughly $0.1 \times R_i$ false positives, and hence that in total there are roughly $0.1 \times \sum_i R_i$ false positives among all the results declared significant, which of course implies a 'global' false discovery rate of 0.1. This may seem too good to be true, but (as explained in Subsection 2.1) it is

likely to yield an approximately correct statement under a variety of data-generating models and even for moderate numbers of hypotheses.[5] On the other hand, it must be recognized that in view of the uncertainties alluded to in Subsection 2.1.1 one would often have reservations about claiming such a bound on the global false discovery rate. All things considered, reporting the individual conservative false discovery rates per multiple testing problem is perhaps a fair and transparent procedure; its only disadvantage is that the more sceptical reader may want to bring down the individual bounds in order to secure the global bound, and the consequences of doing this can only be appreciated by examining the lists of rejections and p-values and lopping parts of the 'map of relationships'.

The second point concerns the sequential nature of the multiple testing problems considered at different stages of the analysis: in the third step, the associations between explanatory variables are restricted to the variables selected in the second step, so that, for example, choosing a different conservative false discovery rate in the second step would change the multiple testing problem in the third. (By comparison, the transition from the first to the second step is hardly a problem in this connection, because the selection of phenotypes is comparable to a decision about which variables to sample prior to the study.) This raises the question of whether it is appropriate to bound the false discovery rate separately at each stage, for it is not obvious that the bound claimed in the third step, which is conditional on the results of the second, is valid. Furthermore, one may again ask whether reporting a 'global' false discovery rate would be more appropriate than reporting the various false discovery rates separately. Unfortunately, there are as yet no definitive answers to these questions (but see Yekutieli et al. (2006) and Yekutieli (2008)), so the control of the false discovery rate promised in the third step is not so easy to justify.

Our best argument supporting the control of the false discovery rate in a 'conditional' multiple testing problem is based on a form of the conditionality principle. Testing the associations between each of the response variables and the explanatory variables *and* testing the associations between *all* the explanatory variables does not appear problematic in general if one controls the false discovery rates separately within each multiple testing procedure (as we have argued above). The difficulty arises when instead of testing

[5]Writing S_i for the number of incorrect rejections pertaining to the i-th response variable, we see that the 'global' false discovery proportion is a random convex linear combination of the individual false discovery proportions which, asymptotically, is bounded by a convex linear combination of their individual bounds: $\frac{\sum_j S_j}{\sum_i R_i} = \sum_j \frac{S_j}{R_j} \frac{R_j}{\sum_i R_i} \lesssim 0.1$ with high probability provided $\frac{S_j}{R_j} \lesssim 0.1$ for all j with high probability.

DOI: 10.2202/1544-6115.1420

the associations between all the explanatory variables one tests the associations between a *selection* of explanatory variables, the selection being based on the first multiple testing procedure: conditionally on the selection, the tests statistics used to test the association of pairs of explanatory variables could be biased (the distribution of the p-values under the null might not be approximately uniform anymore, for example), and consequently bias the second multiple testing procedure (irrespective of whether this is based on Benjamini-Hochberg's, Bonferroni's or any other method). This difficulty is actually quite common; it appears whenever one carries out two statistical procedures sequentially, the result of the first influencing the set-up (but not necessarily the conclusion) of the second, and the graceful way out of it is to invoke the conditionality principle, which in our case would amount to assuming that the distribution of the tests statistics under the null remains approximately uniform conditionally on the selection. However 'obvious' or 'intuitive' the validity of this principle may seem, its application in data analyses is often more a question of faith or philosophy than mathematics, and it is possible to construct data-generating models that violate it (e.g. Helland (1995)), so it must be admitted that the control of the false discovery rate (or indeed of any error rate based on p-values) claimed for the third step should be taken with reservation. Our *feeling* is that the conditionality principle holds approximately in data analyses like the one presented in Section 3.

3 An example: detection of associations between risk measures of heart disease and genetic polymorphisms

The REGRESS dataset (Jukema et al. (1995)) consists of measurements of phenotypic and genotypic variables on 884 male patients suffering from coronary heart disease and having normal to moderately elevated cholesterol levels; according to Maat et al. (1998), it "represents the majority of cardiac patients seen in clinical practice". The **phenotypic variables** include HDL and LDL cholesterol levels, 'family history' (which records whether or not a patient's family has a history of heart disease), age, BMI, and other risk measures of heart disease. The **genotypic variables** consist of the genotypes of about 140 *single nucleotide polymorphisms* (SNPs), each of which is characterized as 'wild type', 'heterozygote' or 'homozygote' depending on

[6] A list of the SNPs (with a few question marks) is given in the Appendix; a file with the REGRESS data can be obtained from A.H. Zwinderman.

the number—0, 1 or 2—of minor frequency or **rare alleles** (the alleles that are less common in the general population).[6] Our objective is to detect possible associations between some of the phenotypic variables and the gene polymorphisms in a **target population** that can be roughly defined as the population of male "cardiac patients seen in clinical practice" in the Netherlands or more strictly as the population of patients satisfying the inclusion criteria set in the REGRESS study.

The phenotypic and genotypic variables considered here have already been the subject of many association studies; see Winkelmann et al. (2000) and Kullo and Ding (2007) for reviews and relevant literature One of our aims is to show that most of the associations established in the literature can be detected using a simple and robust approach in a single scrutiny of the REGRESS dataset. The end product of our analysis consists of a 'map of relationships' summarizing the main findings and which will hopefully be useful in generating working hypotheses and setting up new studies, and ultimately in helping to pinpoint those gene polymorphisms which influence the risk of heart disease in similar populations of patients.

The presentation follows the four steps outlined in Section 2.

3.1 Phenotypic variables

We first examine the 20 phenotypic variables that make up most of the baseline measurements of the REGRESS dataset. Figure 1 shows a dendogram obtained by clustering the 20 variables by means of an agglomerative hierarchical method (implemented in the `R` routine `agnes`) where the 'dissimilarity' between two variables is given by the tenth root of the p-value of a test of association between them and the distance between two clusters equals the average of the dissimilarities between the members of one cluster and the members of the other. Since the p-values are all based on approximately the same sample size, their relative closeness to zero should be a good measure of the strength of the association between two variables; we chose its tenth root rather than the p-value itself for purposes of visualization.

The variables related to lipids, namely **HDL**, **LDL** and **total cholesterol** (the sum of HDL, LDL and VLDL (very low density lipoprotein)), **triglycerides**, as well as **glucose** and the variables measuring the *amount* of proteins are all measured in $mmol/l$ (millimoles per litre); variables measuring the *activity* of proteins are measured in μ/ml; levels of **apolipoprotein(a)** and **fibrinogen** are measured in g/l. **CETP** (cholesteryl ester transfer protein), apolipoprotein(a) and **LPL** (lipoprotein lipase) are known to be involved in the synthesis or metabolism of lipids; fibrinogen and **C-reactive**

DOI: 10.2202/1544-6115.1420

protein are thought to be involved in inflammatory processes associated with arterial damage. All these proteins are thought to be regulated in part by polymorphisms (in homonymous genes) represented in the REGRESS dataset. **MSD** (mean segment diameter) and **MOD** (minimum obstruction diameter), measured in mm, are two variables that quantify the diameter of *unobstruction* in certain segments of the coronary tree. **Stenosis** is the average percentage of constriction over selected blood vessels, and **ejection fraction** is the percentage of blood pumped out of the left ventricle per heartbeat. Finally, **systolic** and **diastolic** blood pressures are measured in $mmHg$ (millimetres of mercury), and **BMI** in kg/m^2.

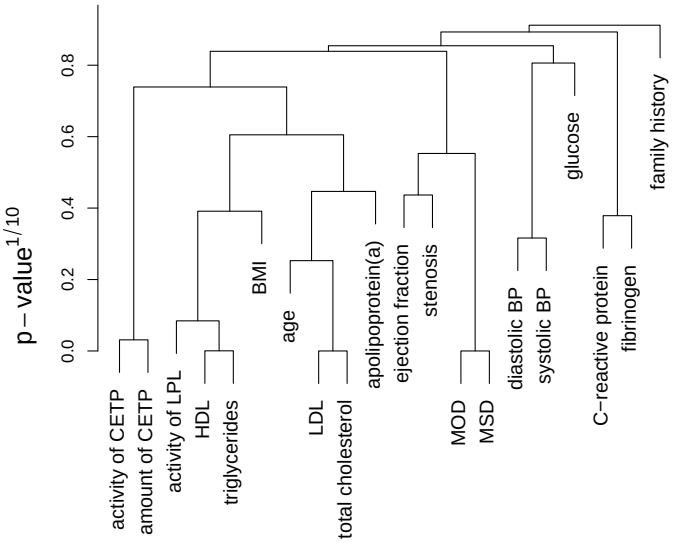

Figure 1: Dendogram summarizing the relationships between the 20 phenotypic variables.

The dendogram provides a simple and sensible picture of the relationships between the phenotypic variables. Since 0.6 in the tenth root scale corresponds to a p-value of about 0.006, we see that the evidence for the associations is quite strong. Unsurprisingly, the evidence for associations between LDL and total cholesterol, HDL and triglycerides, amount of CETP and activity of CETP, MSD and MOD, and systolic BP and diastolic BP is particularly strong. There is also good evidence that both HDL and triglycerides are associated with the activity of LPL, the key enzyme involved in the clearance of triglycerides from plasma. Higher up in the dendogram we find evidence of the associations between the two groups of lipids headed by LDL

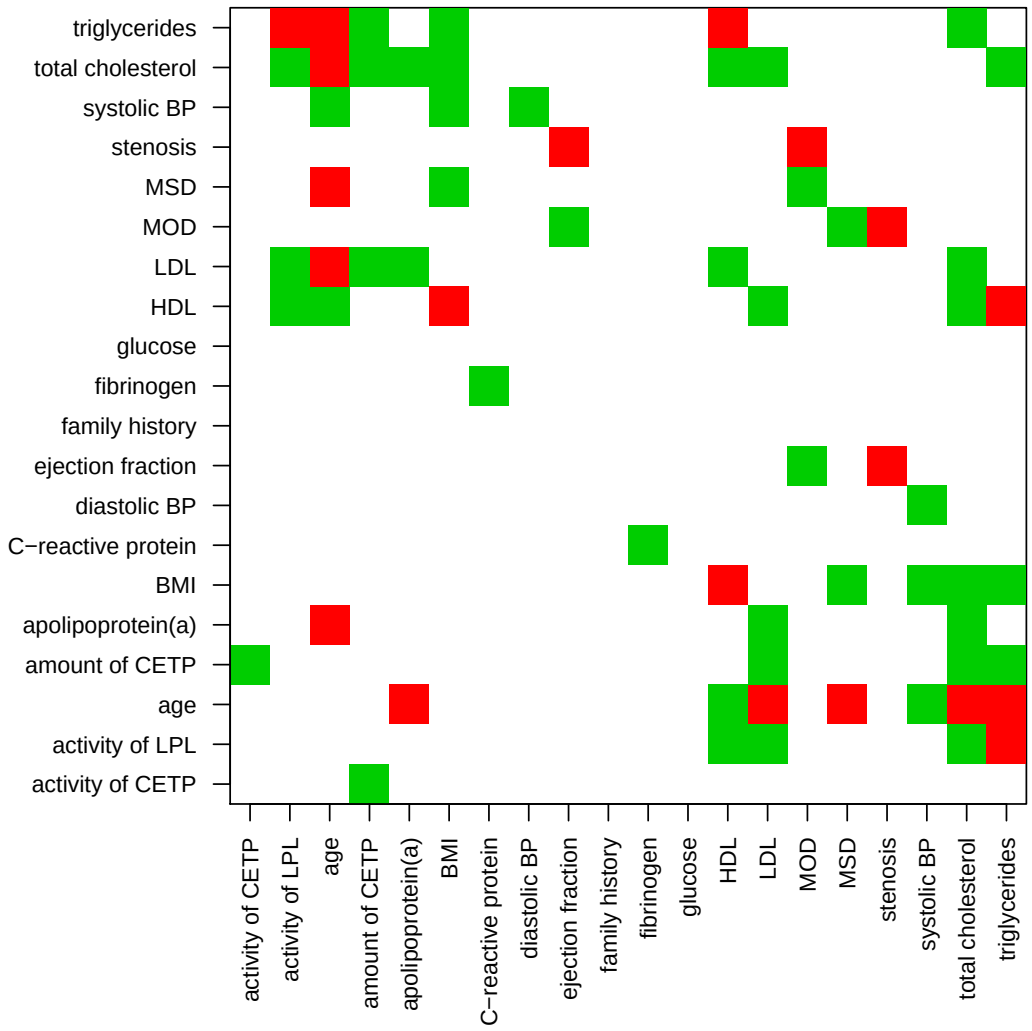

Figure 2: Two-dimensional colour plot characterizing the direction of the 'top' associations—those with p-values below 0.005—between some of the 20 phenotypic variables: positive associations are represented in green, negative associations in red.

and HDL and between these and age and BMI; similarly, we find evidence of the association between MSD/MOD and ejection fraction/stenosis, all of which are angiographic parameters. High levels of fibrinogen and high activity of C-reactive protein have each been associated with heart disease; that they both participate in inflammatory processes and appear to be associated with each other here does of course not mean that there is a causal relation between them, and the same applies to all other associations in the graph.

DOI: 10.2202/1544-6115.1420

Figure 2 gives an overview of the sign or direction of the 0.5% 'top' associations between the phenotypical variables, as determined by the *sign* of the sample correlation coefficient.

Figures 3 and 4 illustrate some of the associations. Despite the strong evidence in favour of these associations, some of them are rather noisy or even difficult to distinguish, such as those between LDL and apolipoprotein(a) levels and between C-reactive protein activity and fibrinogen levels. On the other hand, it is apparent that some of the variables may serve as proxy measures of others (e.g. HDL of triglycerides, MSD of MOD, MOD of stenosis, CETP amount of CETP activity, systolic pressure of diastolic pressure).

The $190 = 20 \times 19/2$ p-values used to construct the dendogram come from Spearman tests (in the case of two continuous variables), Kruskal-Wallis tests (in the case of one continuous variable and one categorical variable), and chi-square tests (in the case of two categorical variables). Besides **family history** of heart disease, the other two categorical variables are systolic BP and diastolic BP; these variables can only be measured in a very discrete manner, but they are also sparsely distributed, so in order to test their association we have categorized them into three classes each: $(0, 120]$, $(120, 140]$, $[140, \infty)$ in the case of systolic BP, $(0, 80]$, $(80, 95]$, $[95, \infty)$ in the case of diastolic BP. All the p-values have been computed with the statistical software `R` and are mostly based on approximate or asymptotic distributions, which is alright since our interest here is on ranking the associations in terms of their strength rather than detecting associations at a given false discovery rate.

The histogram of the 190 p-values, shown in the left panel of Figure 5, is typical of multiple testing problems with a **proportion of true null hypotheses**, γ (recall Subsection 2.1), substantially smaller than 1: a roughly decreasing density with a pronounced peak near zero. As mentioned in Subsection 2.1.1, the height of the histogram near 1 can be taken as an overestimate of γ; thus the histogram suggests that the number of genuine associations among the 190 possible associations is quite large, which is not very surprising.[7] The other histogram of Figure 5 was obtained by simulating the 190 p-values under the assumption that no associations exist. More precisely, we simulated 17 random samples from normal distributions with the same means and variances as the samples of continuous phenotypic variables and three random samples from discrete distributions determined by the frequencies of the categorical variables (all samples being independently generated and having the same missing values as the real data) and then carried out the 190 tests and constructed the histogram of the corresponding p-values. As expected—and in spite of the dependence between the p-values (due to the fact that two p-values can be partly

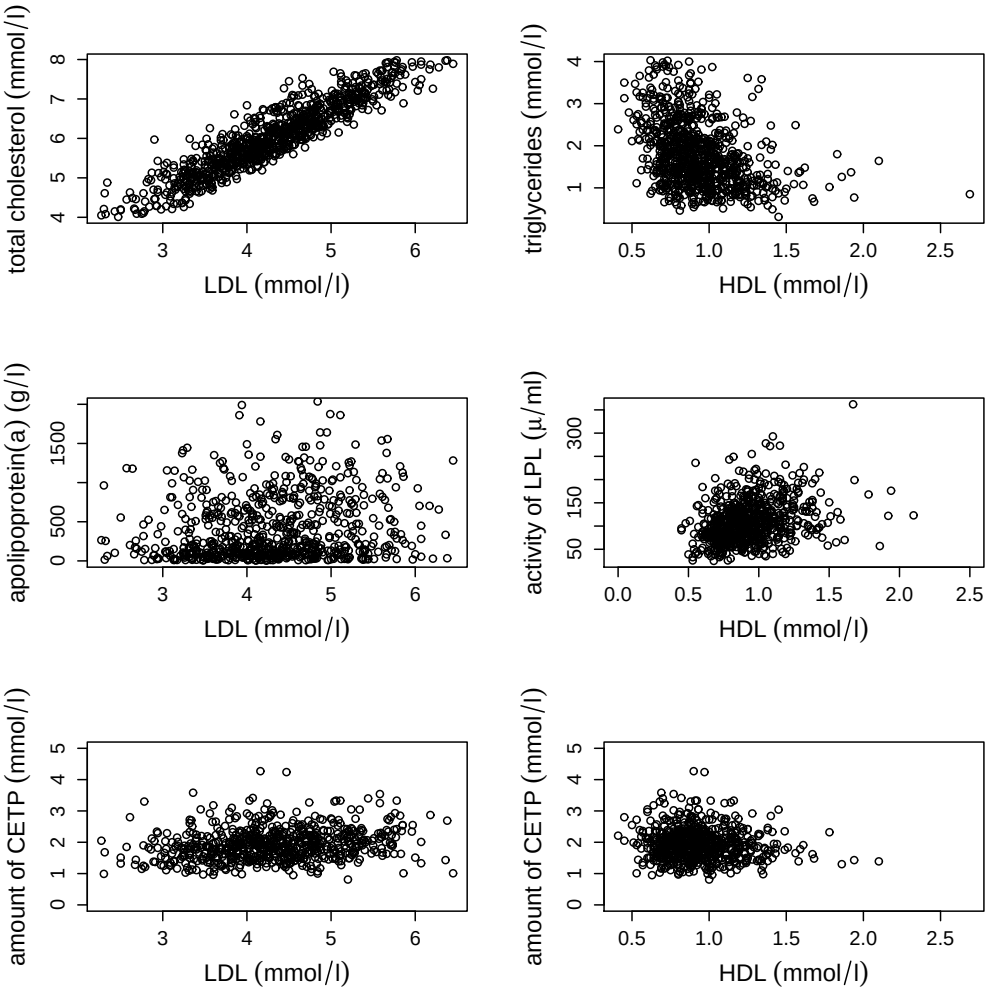

Figure 3: Scatter plots illustrating the associations between some of the lipid-related variables.

based on the same sample)—the histogram of the p-values under the 'null hypothesis' that no associations exist is close to the uniform density.

The objective of this section is really to gain an overall picture of the phenotypic variables and to choose some of these variables on which to concentrate during the rest of the analysis. The **main risk measures of heart disease** are cholesterol levels, hypertension, diabetes, family history of heart disease, BMI, age, and smoking (Winkelmann et al. (2000)), and only the first four are likely to be strongly influenced by the relatively small set of genes considered here. It is thus clear that our association study should fo-

DOI: 10.2202/1544-6115.1420

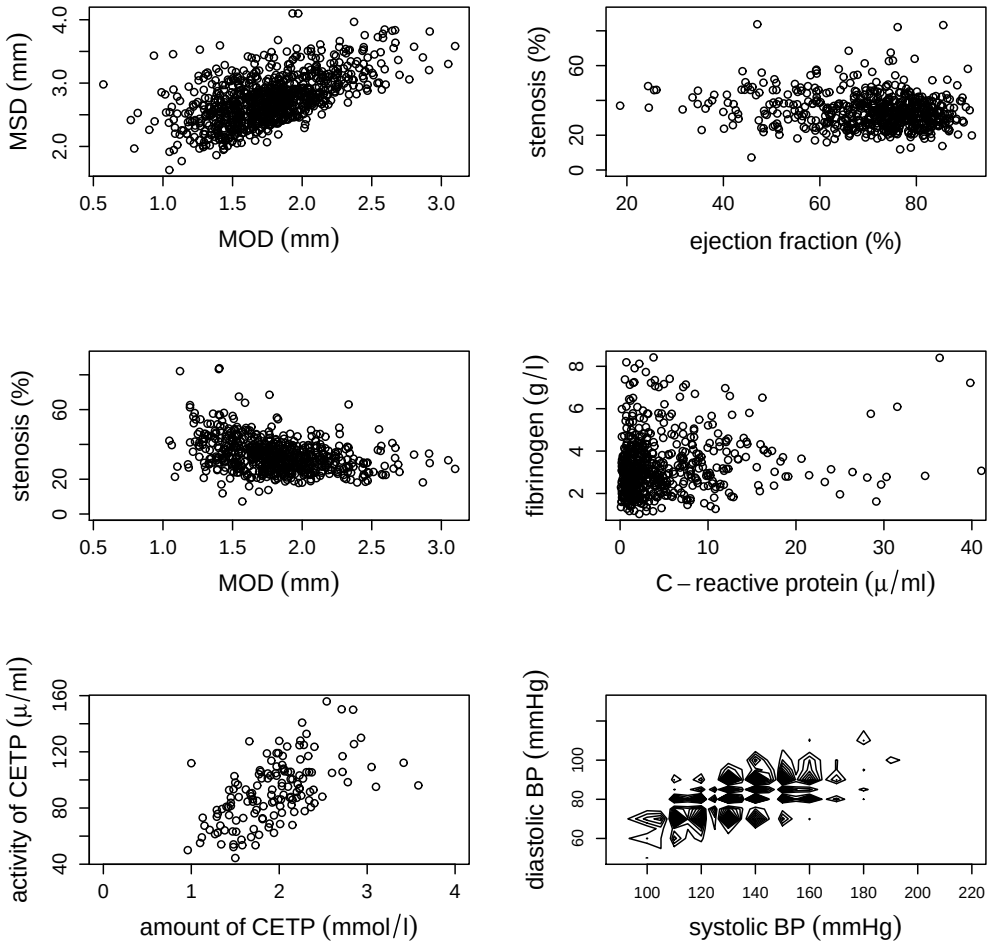

Figure 4: Scatter plots illustrating the associations between some of the angiographic variables, between fibrinogen levels and C-reactive protein activity, between amount and activity of CETP, and between systolic and diastolic pressure.

cus on lipids (HDL, LDL, triglycerides, etc.), blood pressure, glucose and family history. Angiographic parameters like MSD, MOD and stenosis are both measures of heart disease and measures of *risk* of heart disease, and they may also be directly or indirectly influenced by genes. In principle, the activity and/or amount of proteins like apolipoprotein(a), CETP and LPL are interesting variables to consider in parallel with lipids, and fibrinogen levels and C-reactive protein activity are interesting on their own. In view of these observations, in the remainder of our work we shall concentrate on possible associations between the more than 130 SNPs represented in the REGRESS

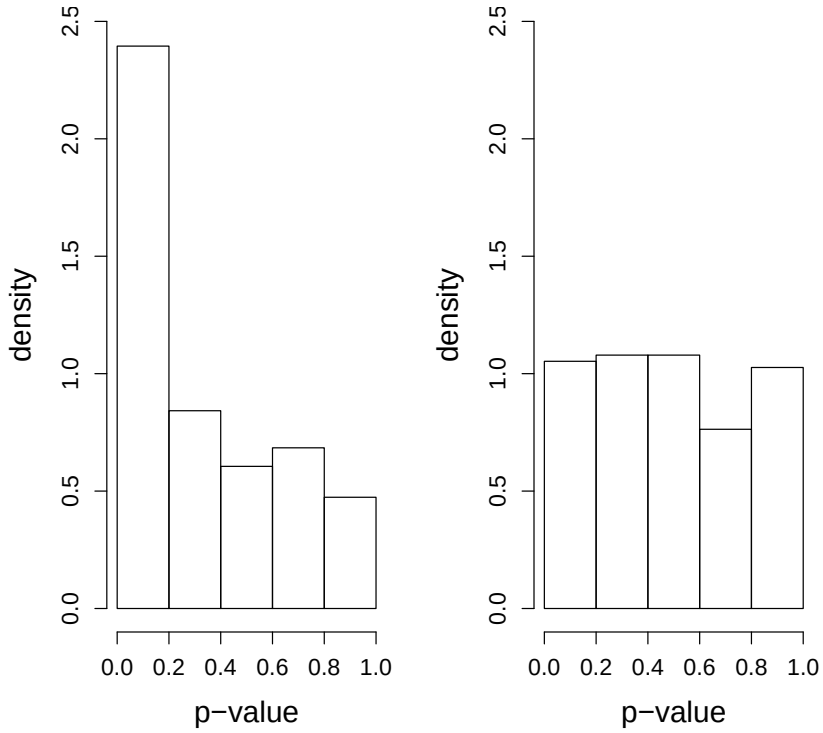

Figure 5: Histograms of the p-values from 190 tests of pairwise independence between 20 variables. *Left panel*: based on the real phenotypic data. *Right panel*: based on simulated data under the 'null hypothesis' that no associations exist.

dataset and each of the following eleven phenotypical variables: HDL, LDL, MSD, systolic BP, glucose, family history, apolipoprotein(a) levels, LPL activity, amount of CETP, C-reactive protein activity, and fibrinogen levels. The reasons for including only one of the blood pressures, only one of the angiographic measures, and only one of the CETP variables are obvious; the lack of a clear association between LDL and the levels of apolipoprotein(a) (see Figure 4), despite the result of the test, seems a good reason for looking separately at this protein.

[7] All histograms in this work are constructed with Scott's rule as implemented in `R`; with uniform-like densities, it tends to perfom better than Sturges's in terms of the L^1 norm.

DOI: 10.2202/1544-6115.1420

3.2 Associations between phenotypic variables and polymorphisms

In the second step of our analysis we carry out a multiple testing study to detect associations between each of the eleven phenotypic variables selected in the previous section and 134 of the polymorphisms included in the REGRESS dataset (some of the polymorphisms listed in the Appendix have been excluded because they are present in only one form in this population of patients). A polymorphism is characterized by 0, 1 or 2—or, in the case of rarer alleles, by 0 or 1—rare alleles. Whatever the case, a test of independence between a phenotypic variable and a given polymorphism amounts to a test of the hypothesis that the distribution of the former is the same across the two or three populations determined by the latter. Consequently, the Scholz and Stephens (1987) test, which is based on empirical distributions (on ranks, actually) and can be used to test the equality of two or more populations whose distributions are continuous, and the exact (conditionally on the marginal frequencies) chi-square test of independence, seem to be the ideal tools for our purposes since they are powerful as well as consistent against all departures from homogeneity/independence.

As mentioned earlier, the main difficulty in performing multiple testing studies based on non-parametric tests lies in the accurate computation of p-values. In contrast with the analysis of the preceding subsection, the use of asymptotic null distributions will not do here because in most cases one of the populations (e.g. that corresponding to homozygotes) will have very small numbers. The only solution, therefore, is to compute the p-values by simulation, which can be a rather demanding task. For example, to compute (with high probability) a p-value of the Scholz-Stephens statistic correctly up to four decimal places requires at least 10 million simulation runs, which can take five days of computing time on a PC with a CPU of 3.20 GHz and RAM of 2GB. Fortunately, only a few very small or large p-values need to be computed with this accuracy; the rest can be computed to two decimal places without affecting the result of the multiple testing procedure.

The p-values of the Scholz-Stephens test presented below were computed in `R` with Fritz Scholz's package `adk.test`, and they are based on the permutation distribution rather than on the null distribution in order to deal appropriately with ties. The p-values are generally similar to those estimated by simulation under the null hypothesis, except in the case of HDL and glucose, whose samples have many ties; in the latter case the data have only one decimal place and contain only 70 different values, yielding a seriously distorted histogram of p-values calculated assuming no ties. The results based

Phenotypic variable	Polymorphisms (ranked by p-value)	Sign	P-value	Guess at $\bar{\gamma}$	Expected no. false positives
HDL	*CETP (-629)C>A*	+	≈ 0	0.8	1
	CETP TaqIB	+	0.00001		
	CETP (-1337)C>T	+	0.00040		
	CETP 784CCC>A	+	0.00048		
	CETP (-2708)G>A	+	0.00131		
	CETP MspI	−	0.00172		
	LIPC (-710)T>C	+	0.00212		
	LPL HIND3	+	0.00455		
	LIPC (-514)C>T	+	0.00948		
	APOE C112R	−	0.00948		
LDL	None	+	...	0.8	...
MSD	*ABCA1 InsG141*	+	0.00026	≈ 1	0.07
	ABCA1 UTR1395	+	0.00041		
	LPL Ser447Stop	+	0.00404*		0.54*
Systolic BP	None	...	...	≈ 1	...
Glucose	*LPL Asn291Ser*	−	0.00006	≈ 1	0.008
Family history	*PAI-1 (4G5G)*	−	0.0027252	≈ 1	0.36
Apolipoprotein(a) levels	*GNβ3 C825T*	+	0.00435	≈ 1	0.59
	NOD2 InsG3020	−	0.00442		
LPL activity	*LPL HIND3*	+	0.00080	≈ 1	0.11
Amount of CETP	*CETP (-2708)G>A*	−	≈ 0	≈ 1	1.73
	CETP TaqIB	−	≈ 0		
	CETP MspI	+	≈ 0		
	CETP (-629)C>A	−	≈ 0		
	CETP (-1337)C>T	−	≈ 0		
	CETP 784CCC>A	−	≈ 0		
	CETP (-972)G>A	−	0.00029		
	eNOS G894T	−	0.00742		
	CETP IIe405Val	−	0.01288		
C-reactive protein activity	*APOE C112R*	−	0.00091	≈ 1	0.12
	CETP MspI	−	0.00436*		
	SCARB1	−	0.00637*		0.85*
Fibrinogen levels	None	...	...	≈ 1	...

Table 1: Results of the multiple testing procedures aimed at detecting associations between each of 11 phenotypic variables and 134 genetic polymorphisms. The bounds on the expected numbers of false positives marked by an asterisk refer to the selection of *all* polymorphisms indicated; the others refer to the selection of all polymorphisms indicated except those possibly marked by an asterisk.

DOI: 10.2202/1544-6115.1420

on the chi-square test—namely those pertaining to systolic BP and family history—were obtained in `R` with the routine `chisq.test`, which estimates p-values by simulating contingency tables under independence conditionally on the marginals.

Table 1 summarizes the results of the multiple testing procedures, namely the 21 polymorphisms that are putatively associated with a phenotypic variable, the sign of the association (positive or negative as determined by the sign of the sample correlation coefficient), the corresponding p-values, the estimated false discovery rate incurred in each case, and a rough, conservative estimate of the proportion γ of true null hypotheses (in this case the proportion of polymorphisms that are truly independent of the relevant phenotypic variable), which is denoted by $\bar{\gamma}$. In most cases, the bound on the expected number of false positives corresponds to a false discovery rate of 0.10 or less; the case of family history (one rejection at a false disvovery rate of 0.36) and that of apolipoprotein(a) (two rejections at a false disvovery rate of 0.28) are exceptions; the other exceptions are indicated by an asterisk. We shall now explain the results of HDL and LDL in some detail.

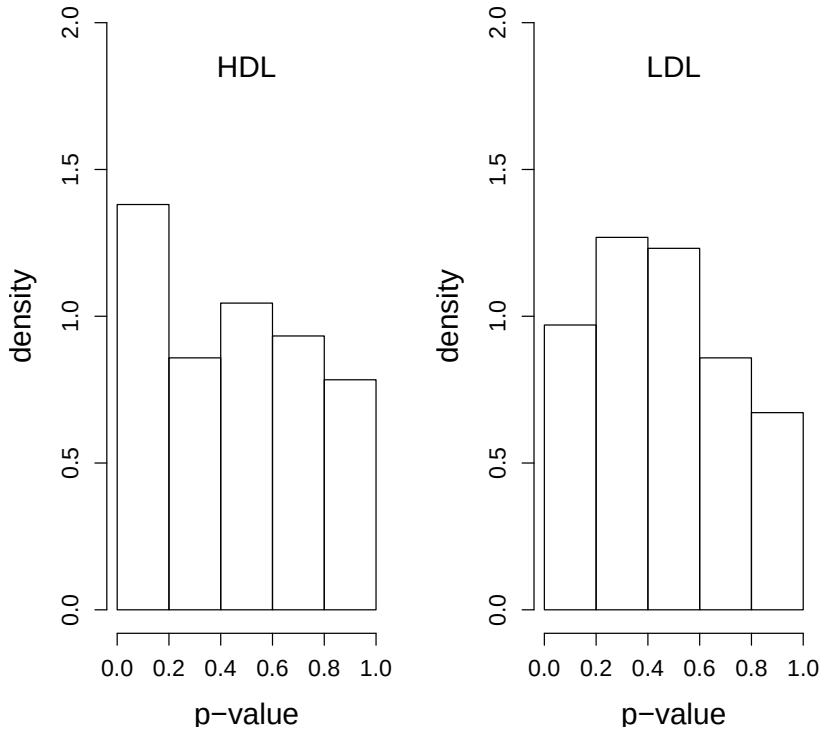

Figure 6: Histograms of the p-values obtained from 134 tests of pairwise independence between each of HDL and LDL and 134 polymorphisms.

Figure 6 shows the histograms of the p-values obtained from the tests of independence between each of HDL and LDL and 134 polymorphisms. The histogram on the left is typical of data with some 'signal': a roughly decreasing function with a peak near zero and going below 1 in the left neighbourhood of 1. This last feature suggests (recall Subsection 2.1.1) that the fraction of polymorphisms that are truly independent of HDL is as small as 0.8, or, in other words, that HDL depends on 27 or more polymorphisms.

The 'top' (i.e. smaller) 10 p-values of HDL range from nearly 0—that of the *CETP (-629)C>A* polymorphism in the CETP gene—to 0.00948—that of *APOE C112R*. The false discovery rate incurred by declaring the corresponding polymorphisms to be associated with HDL is therefore estimated as $0.00948 \times 134/10 \approx 0.127$. However, taking 0.8 as an upper bound on γ (the proportion of polymorphisms that are truly independent of HDL) the false discovery rate can be estimated somewhat less conservatively as $0.8 \times 0.00948 \times 134/10 \approx 0.101$, which

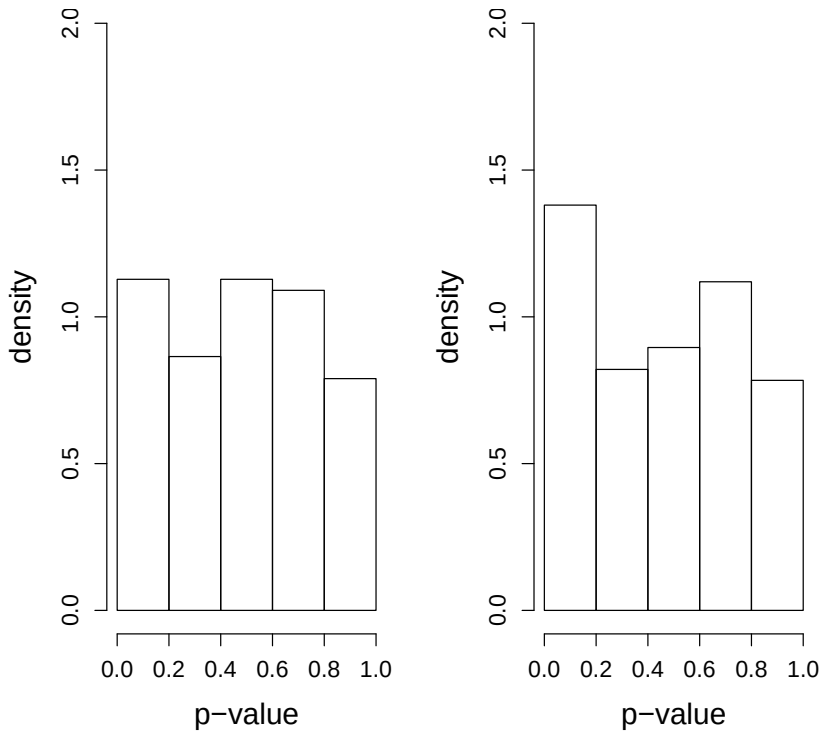

Figure 7: Histograms of p-values resulting from 134 tests of independence between HDL and the polymorphisms based on two randomly chosen halves of the REGRESS dataset.

DOI: 10.2202/1544-6115.1420

can be interpreted as saying that among the 10 polymorphisms thus singled out at most one is expected to be spurious. For comparison, we note that a Bonferroni procedure of level 0.1 would reject the p-values below $0.000746 = 0.1/134$, which would amount to selecting only the top four

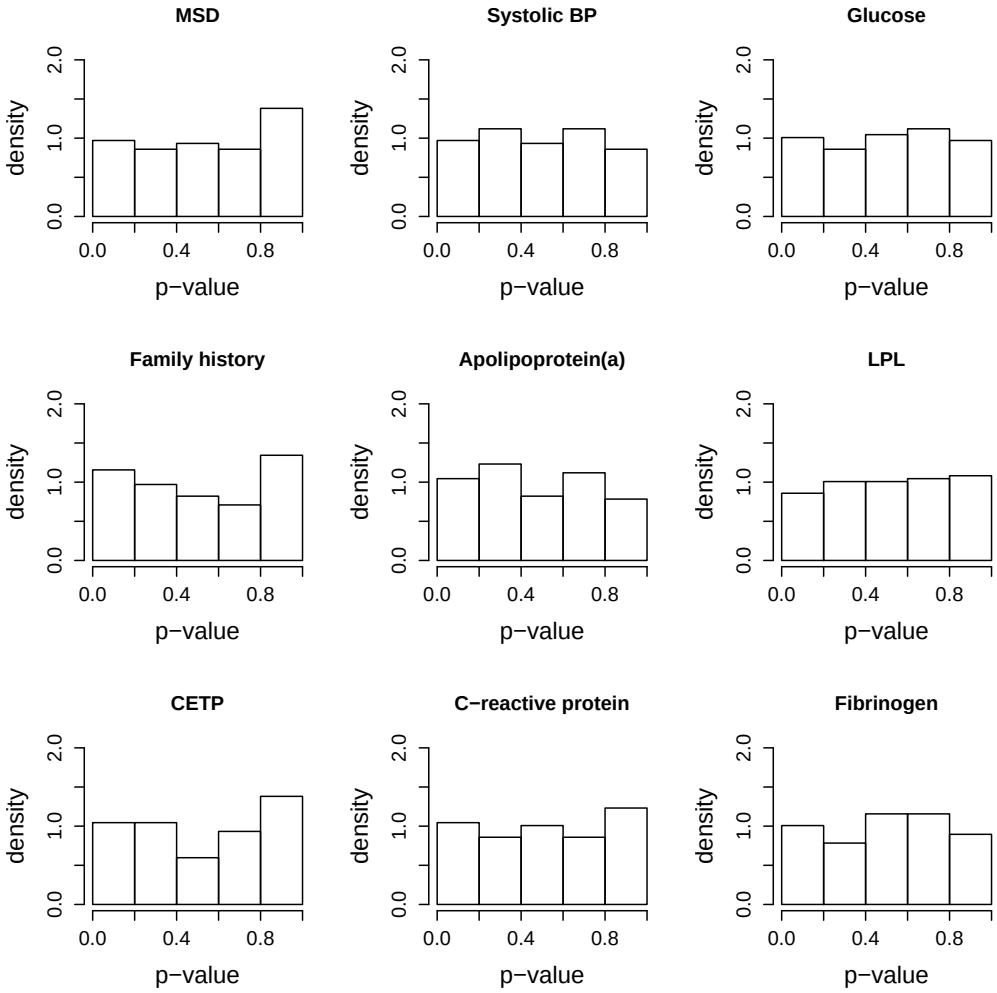

Figure 8: Histograms of the p-values obtained from tests of association between each of several phenotypic variables and 134 polymorphisms.

associations. On the other hand, assuming that there could be as many as 27 genuine associations, the 10 selected by the Benjamini-Hochberg method could represent only about 37% of the 'discoveries to be made'.

The upper bound of 0.8 on γ we have assumed here should be thought of as intrinsic to this population of patients and to the variables in question (HDL and the 134 polymorphisms): it is expected to be more or less insensitive to the particular sample drawn as well as to sample size. To see how sensible this assumption is, we have randomly split the REGRESS dataset into two subsets of 442 patients and applied the multiple testing procedure on HDL to each of them. The resulting histograms of p-values, shown in Figure 7, though less peaked near zero than the first histogram in the left panel of Figure 6 (the tests based on 442 patients being necessarily less powerful than those based on 884 patients), look roughly similar to each other, especially near 0 and 1. This suggests that the empirical distribution of the p-values is more or less invariant with respect to the drawing of samples of the same size and seems to support the assumption that $\gamma \leq 0.8$.

As seen in the third column of the table, most polymorphisms are positively associated with HDL, which is to say that the bigger the number of rare alleles the bigger the levels of HDL tend to be; exceptions are the *CETP MspI* and *APOE C*112*R* polymorphisms.

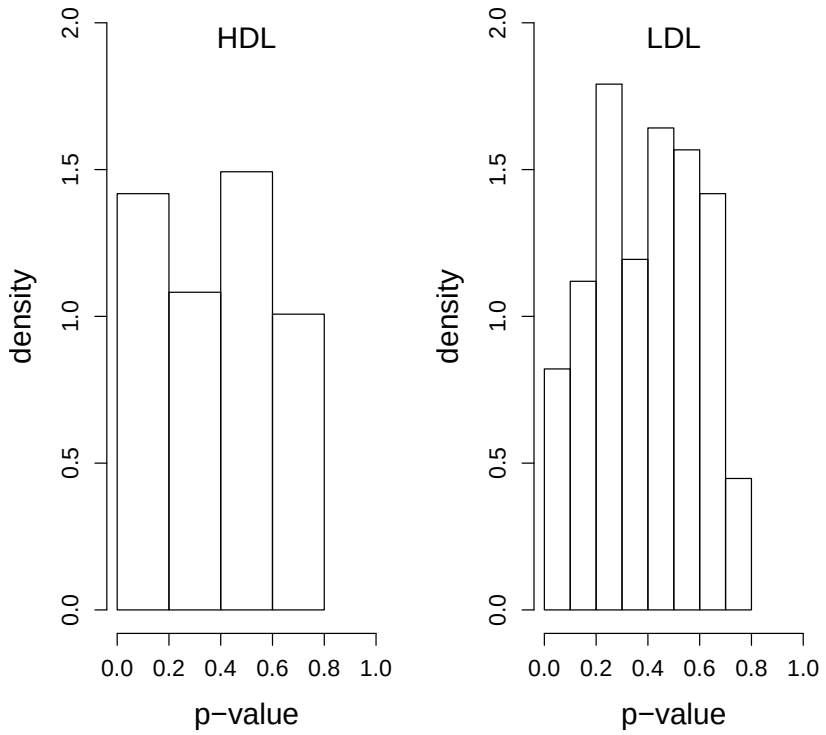

Figure 9: Histograms of the *approximate* p-values obtained from 134 tests of association between each of HDL and LDL and 134 polymorphisms; compare with the corresponding histograms of Figure 6.

DOI: 10.2202/1544-6115.1420

In the case of LDL, the histogram of the 134 p-values (Figure 6) deviates somewhat from the uniform pattern, but roughly in the middle of the $[0, 1]$ interval rather than near zero. Using the Benjamini-Hochberg method with varying q we get no rejections for values of q in $[0, 0.7]$, while at $q = 0.8$ we suddenly get 40% rejections, which obviously entails an unaffordable number of false positives. Since the smallest p-value is 0.01957, it is also clear that no Bonferroni procedure will select a single polymorphism in this case.

Figure 8 shows the histograms of the p-values obtained from the multiple testing procedures pertaining to the other nine phenotypic variables. The histograms look rather flat and thus support the basic assumption behind the Benjamini-Hochberg method: that the empirical distribution of the p-values computed under the null hypotheses is close to the standard uniform distribution. Accordingly, the overestimates of γ given in Table 1, obtained by visual inspection of the histograms of p-values, suggest that HDL and LDL are the only phenotypes that depend on a substantial number of polymorphisms.

On the other hand, p-values computed with the incorrect null distribution can yield histograms that deviate more seriously from the uniform density, possibly entailing unreliable bounds on the false discovery rate. To illustrate this point we have used the `adk` package to compute *approximately* the p-values pertaining to HDL and LDL and compared their histograms, shown in Figure 9, with those obtained with the correct null distribution: although the approximate p-values of HDL yield a histogram rather similar to that of Figure 6, those of LDL suggest a somewhat greater deviation from uniformity; also, none of the histograms has observations in $[0.8, 1]$.

At this stage we could already visualize the findings listed in Table 1, for instance by plotting the distribution of the phenotypic variables across the strata determined by the relevant polymorphisms; however, it will be more economical to postpone this to the 'verification stage' (fourth step) and to concentrate first on the associations between the selected polymorphisms.

3.3 Associations between selected polymorphisms

Some of the 21 polymorphisms listed in Table 1 may be associated with each other, and it is obviously of interest to try to detect and illustrate the associations between them. Some of these associations will not be surprising and may simply reflect linkage or linkage disequilibrium (like those between the polymorphisms of the CETP gene), but others may be more interesting.

Our strategy here is to carry out the 210 ($21 \times 20/2$) tests of pairwise independence between the 21 selected polymorphisms, apply the Benjamini-Hochberg method to the resulting p-values, and then illustrate the associa-

tions selected at a given false discovery rate. The test of independence is the exact chi-square test mentioned at the beginning of the previous subsection.

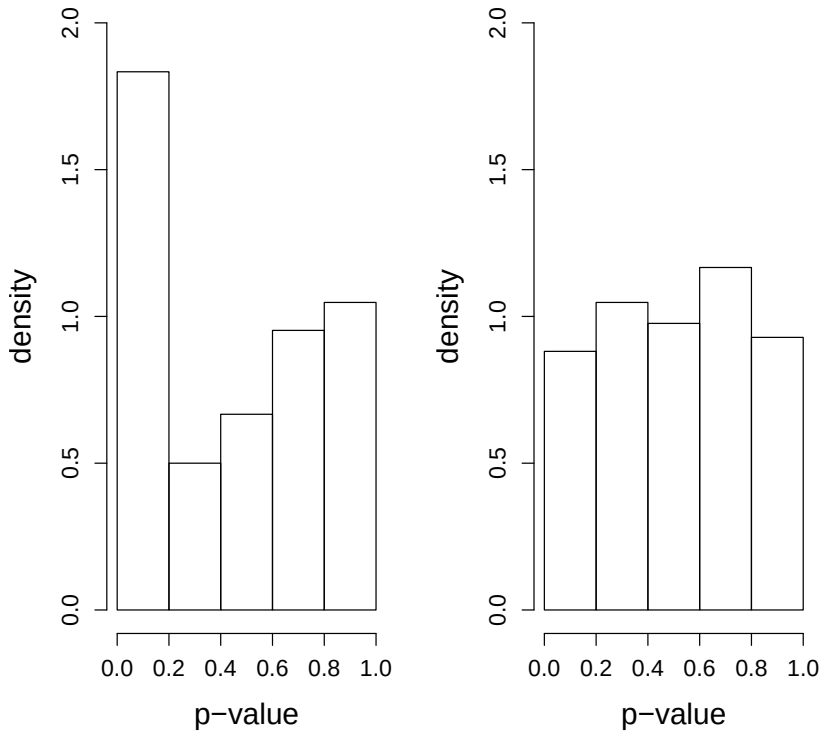

Figure 10: Histograms of p-values resulting from 210 chi-square tests of pairwise independence between the 21 selected polymorphisms: *Left panel*: based on the real data. *Right panel*: based on simulated data under the 'null hypothesis' that no associations exist.

The left panel of Figure 10 shows the histogram of the p-values obtained from the 210 tests of independence. Its height in the left neighbourhood of 1 suggests that there are relatively few associations between polymorphisms. In order to check whether the histogram of the p-values computed under the 'null hypothesis' of no associations is approximately uniform, we have randomly assigned the genotypes of the 21 polymorphisms to the patients according to the estimated frequencies and carried out the chi-square tests; the histograms of the resulting samples of p-values, one of which is shown in the right panel of Figure 10, seem to conform with the expected pattern. (Given the present sample sizes, the distribution of the p-values of the exact chi-square tests is virtually continuous, but, again, approximate uniformity should be checked because the p-values are dependent.)

DOI: 10.2202/1544-6115.1420

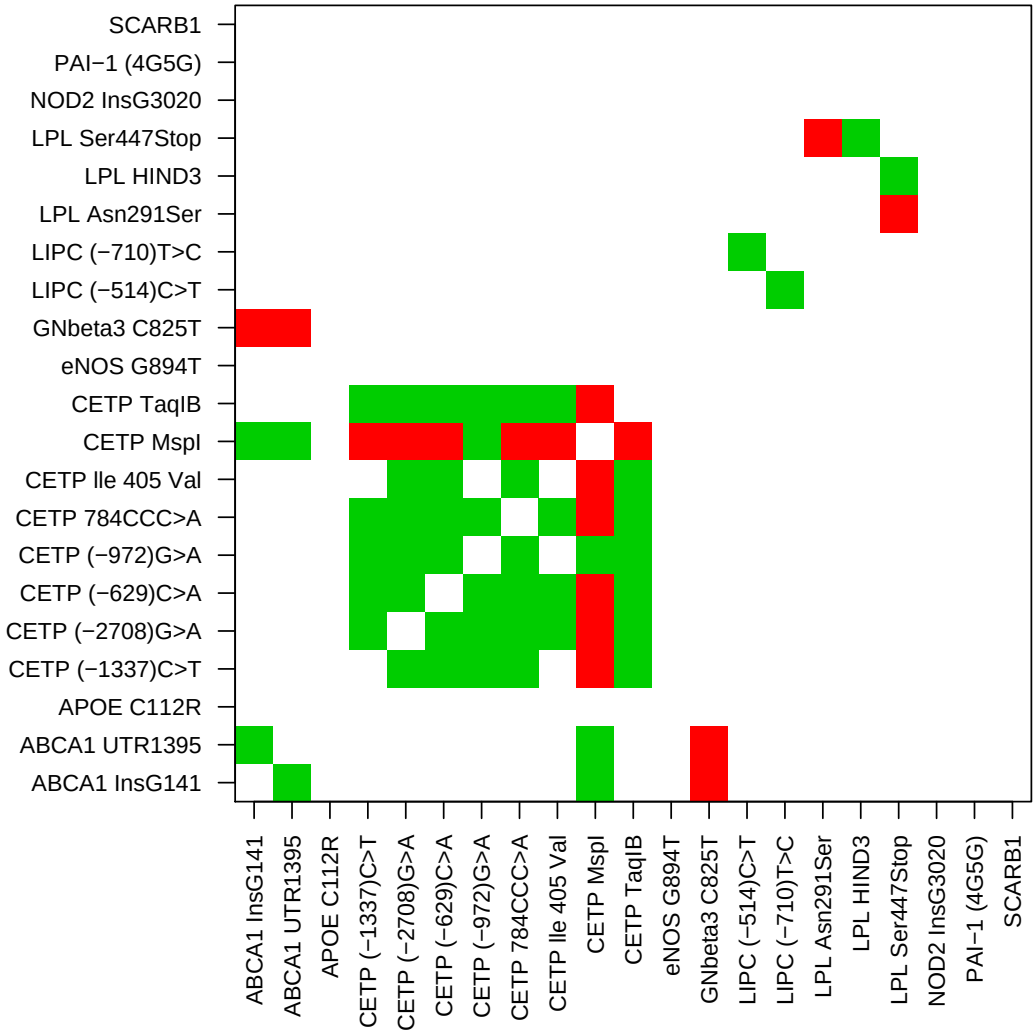

Figure 11: Two-dimensional colour plot characterizing the direction of the 'top' 34 associations singled out at a false discovery rate of 0.10 between the 21 selected polymorphisms: positive associations are indicated in green, negative associations in red.

At a conservative false discovery rate of 0.10 we single out the 'top' 34 associations with estimated p-values ranging from 0.0000001 to 0.0147131; thus, among the 34 associations, we may expect three or four to be spurious. (There is no point in using the adaptive version of the Benjamini-Hochberg here.) Figure 11 gives an overview of the 34 associations between the polymorphisms and indicates whether they are positive or not (according to the sign of the sample correlation coefficient). From the plot we recognize

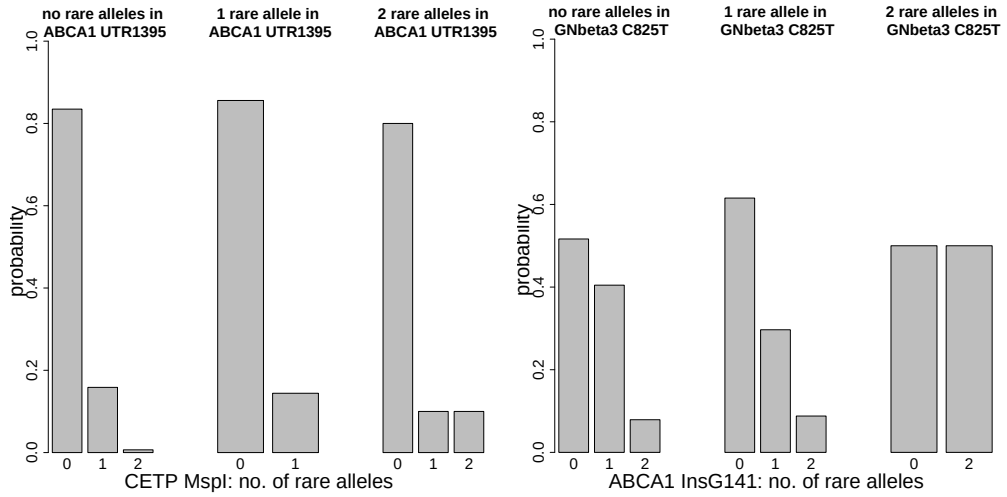

Figure 12: Illustration of two associations between polymorphisms of different genes: histograms of the number of rare alleles in one polymorphism as functions of the number of rare alleles in another.

four clusters of closely associated polymorphisms corresponding to the same gene: eigth of the CETP gene, two of ABCA1, two of LIPC and three of LPL. We note the inverse (or negative) association between *CETP MspI* and most of the other polymorphisms of the CETP gene. Interestingly, we only find evidence of four associations between polymorphisms of different genes: the associations between *CETP MspI* and the two *ABCA1* polymorphisms, and between these two and *GNβ3 C825T*, all of which appear near the bottom of the list (in ranks 30 to 33), with p-values ≥ 0.0097553. Two of these associations are illustrated by the histograms of Figure 12: the first is a positive but apparently weak association, the second an overall negative, not particularly clear association.

The sets of histograms of Figure 13 illustrate other associations. The first exhibits a straighforward relationship that typifies the majority of the associations we have singled out. The second and third exemplify negative, more or less straightforward relationships. The last one serves to show that *CETP MspI* and *CETP (-972)G>A*, despite having respectively negative and positive associations with the other CETP genes, have a roughly positive, but far from straightforward, association with each other.

DOI: 10.2202/1544-6115.1420

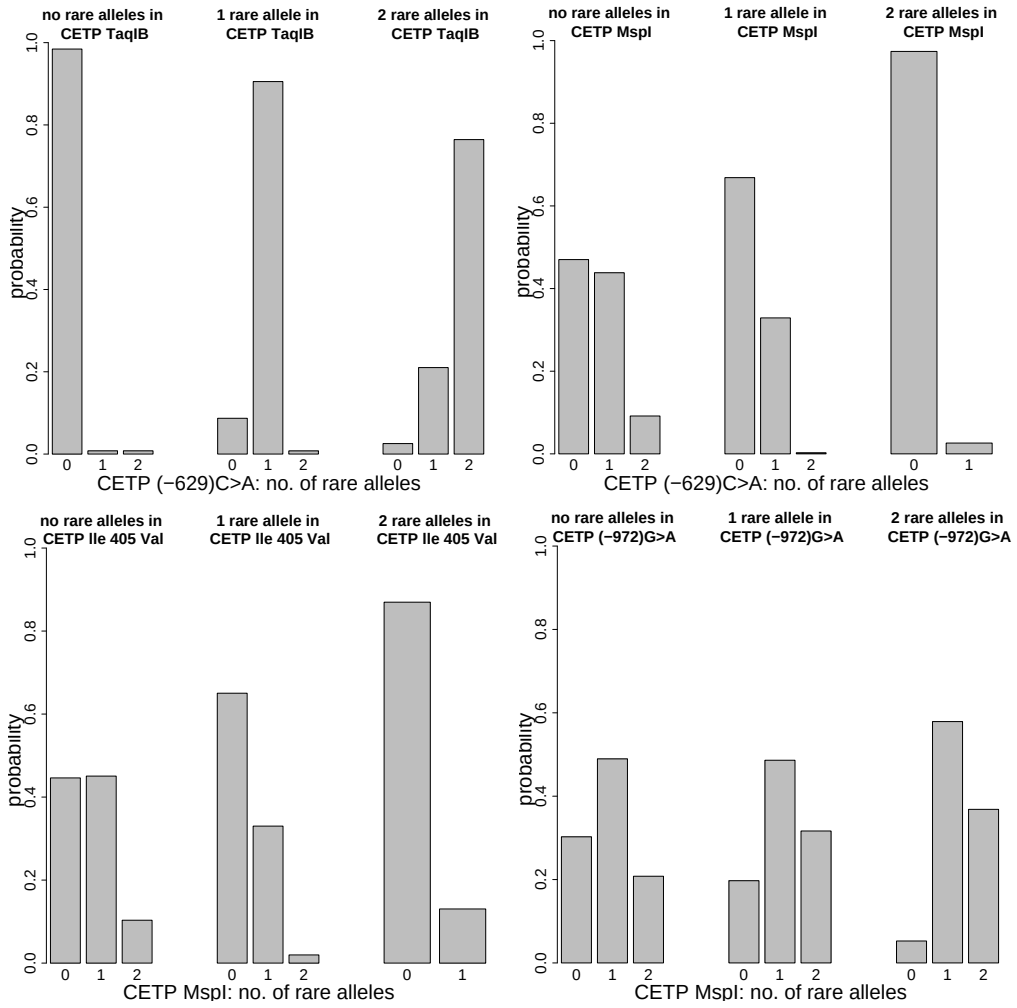

Figure 13: Illustration of some of the associations between polymorphisms: histograms of the number of rare alleles in one polymorphism as functions of the number of rare alleles in another.

3.4 Verification and overview of the findings

The last paragraph is already part of the fourth step and implies that the associations between the selected genotypes have been verified. Indeed, most of these associations are straightforward and, after the fact, rather plausible. In the rest of the paper we shall be concerned with illustrating and verifying the associations between the phenotypes and the 21 polymorphisms by looking at the distribution of a phenotypic variable within the subpopulations with and without rare alleles in the relevant polymorphism and across age

and BMI strata. These are normally regarded as variables one should 'correct for' in epidemiologic studies, and as such they are often used together with the genotypes as explanatory variables in regression models of phenotypic variables as a means of gaining power of detection. Our motivation for considering them here, however, is that certain combinations of age and BMI strata are likely to overlie interesting sub-groups of patients which a *post hoc* analysis may help uncover and whose identification may subsequently be useful in drawing up explanations and in planning more refined studies. For instance, we have already mentioned that the effect of certain genotypes on indicators of heart disease will be more visible in younger than in older patients; but one may also anticipate that the effect of others will be visible mostly in obese individuals. Because this is a rather tedious and lengthy part of the analysis, we shall examine only HDL and MSD in detail and reproduce just a few representative graphs; but the relevant information concerning all the variables will be given at the end in the 'map or relationships'.

HDL appears to be associated with 10 polymorphisms, six of which in the CETP gene. Overall, the direction of the associations (third column of Table 1) is consistent across the strata, but, as expected, the hard evidence in favour of an association usually comes from the more populated strata, namely those of **non-obese**—with BMI in $(18, 30]$—patients; see Figure 14. Thus, in the case of *CETP MspI* the evidence is particularly strong in both **young**—with age in $(30, 55]$—and **old**—with age in $[55, 71)$—non-obese patients, and not particularly visible in **obese**—with BMI in $(30, 40]$—patients, judging by the p-values of the Wilcoxon-Mann-Whitney test comparing the phenotype in the two subpopulations without and with rare alleles, given in brackets above the boxplots together with the sample sizes.[8] Essentially the same applies to *LPL HIND3*, except that the association of *LPL HIND3* with HDL in the stratum of **young and obese** individuals is inverted; according to the p-value this could be due to chance, but an inversion (relative to the direction in the other strata) of the association with HDL in this sub-group is seen in other polymorphisms as well, namely in *CETP (-2708)G>A*, *CETP (-1337)C>T*, *CETP 784CCC>A* and *CETP TaqIB*, which could be a useful observation.

Frequently, an association is visible in only one of the two more populated strata: in the *CETP (-629)C>A*, *CETP (-2708)G>A*, *CETP (-1337)C>T*, *CETP 784CCC>A*, *CETP TaqIB*, *LIPC (-514)C>T* and *LIPC (-710)T>C* polymorphisms, for instance, the association is clear in the stra-

[8]The value of *post hoc* p-values lies not so much in the evidence they provide in favour of an association—which is conditional on having called a result 'significant'—as in their capacity to tell us *where* the association is likely to be present.

DOI: 10.2202/1544-6115.1420

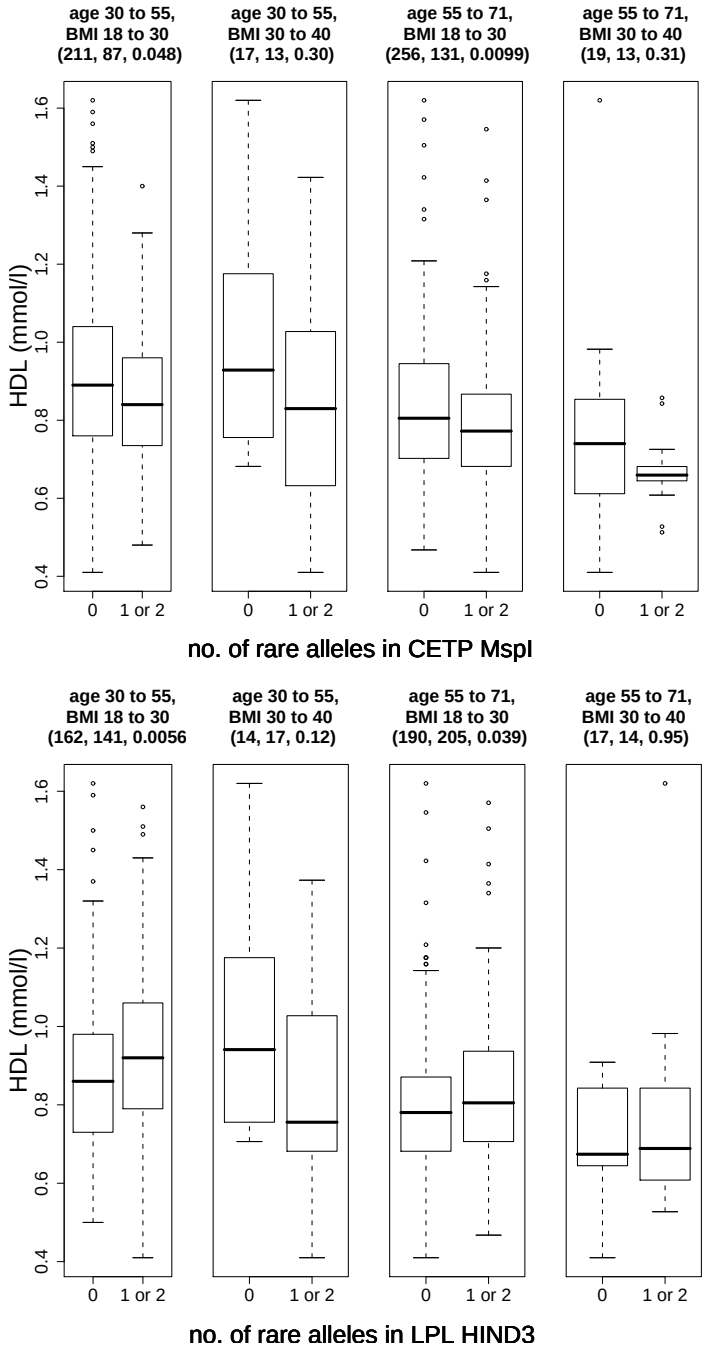

Figure 14: Boxplots of the distribution of HDL by number of rare alleles in the *CETP MspI* and *LPL HIND3* polymorphisms stratified by age and BMI.

tum of **old and non-obese** patients and hardly visible in that of **young and non-obese**. However, it can be difficult to say whether this is due to an imbalance in sample sizes or not. In the case of the *CETP (-629)C>A* polymorphism (Figure 15) we can actually see that this phenomenon cannot be explained by differences in sample size, because the evidence in favour of an association with HDL also comes from the small stratum of **old and obese** patients, which indeed suggests that the association is particularly strong in **older patients**.

Besides *CETP (-629)C>A*, there is another polymorphism for which evidence of an association is found in small strata: *APOE C112R*, whose association with HDL seems to exist only in **young and obese** patients (in young patients at any rate); see Figure 15.

The observations we have made regarding HDL and MSD indicate that in order to appropriately model (e.g. with a view towards testing hypotheses about relevant parameters) the phenotype as a function of the number of rare alleles in a polymorphism one may need to use a (regression) model that accounts for several interactions with age and BMI, since the intensity, and possibly the sign, of the associations may vary according to these variables. More importantly, they suggest that studies intending to go beyond the evidence presented here and to investigate further the role of the selected polymorphisms should focus on more specific patient populations, since, for example, some of the genetic variants we have considered seem to exert an effect on HDL only in the sub-populations of old or of young and obese patients. The same applies to the other phenotypes.

We turn to the associations between **MSD** and the polymorphisms *ABCA1 InsG141*, *ABCA1 UTR1395* and *LPL Ser447Stop*. The first two are positively related (see Figure 11), and their positive association with MSD is very similar, so only that of *ABCA1 InsG141* is illustrated in Figure 16. The effect of these ABCA1 polymorphisms on MSD is seen mainly in the **old, non-obese** group, but the **young, non-obese** group also provides some evidence in the same direction; the inconsistent trends suggested by the other two strata can be attributed to chance. The second plot in Figure 16 suggests that the association between *LPL Ser447Stop* and MSD fails to materialize when viewed against age and BMI. However, by distinguishing between one and two rare alleles and using the Kruskal-Wallis test we see some evidence that the association holds in the **non-obese** group (cf. the last two plots of Figure 16 with the top ones of Figure 17).

The illustration and verification of the other associations goes on very much along these lines, and the best way of presenting it and assimilating it is by means of the map of relationships presented below. Let us just

DOI: 10.2202/1544-6115.1420

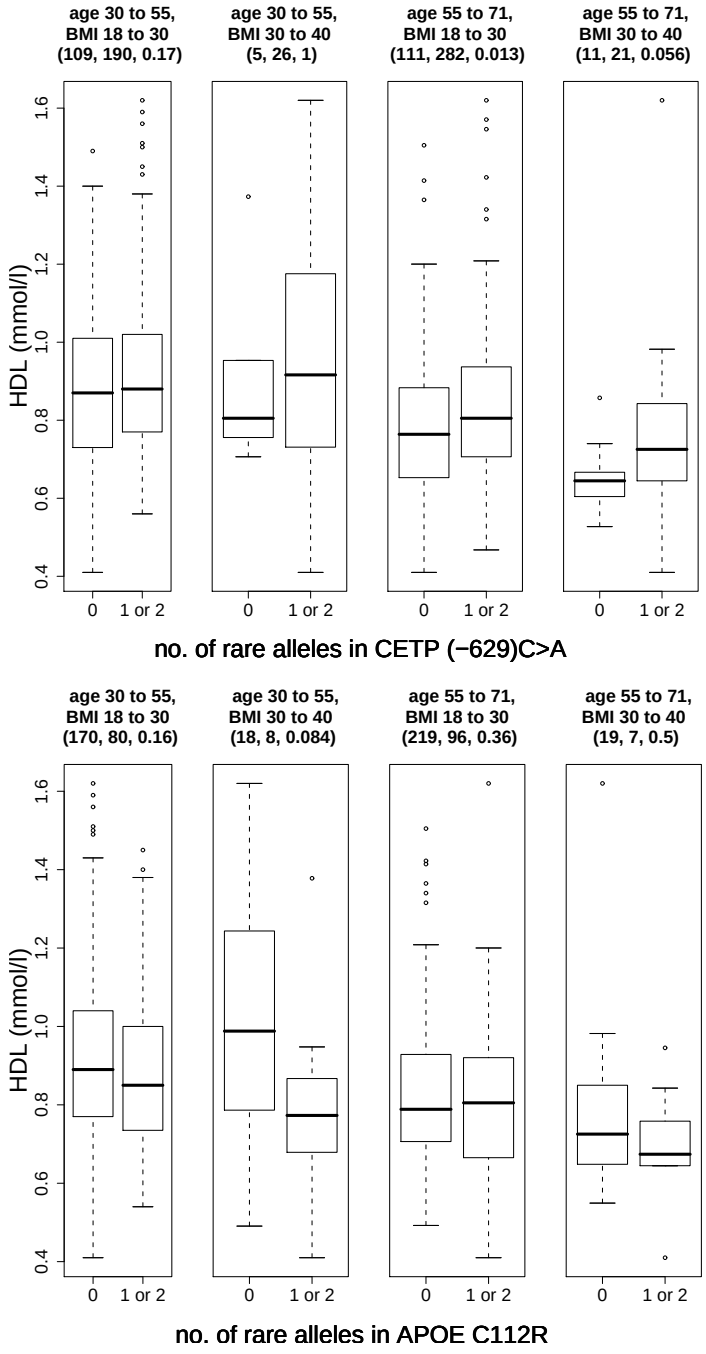

Figure 15: Boxplots of the distribution of HDL by number of rare alleles in the *CETP (-629)C>A* and *APOE C112R* polymorphisms stratified by age and BMI.

mention two more or less special cases: the association between C-reactive protein activity and *SCARB1*, and that between **family history** and *PAI-1 (4G5G)*. The first of these becomes visible—and in the bigger strata—only if one distinguishes between one and two rare alleles; see Figure 17.

The second, illustrated in Figure 18, is visible in the bigger strata and consistent in the obese and non obese groups, but it is difficult to interpret: apparently, the main difference between the group with and without family history is that the former has *fewer* individuals with two rare alleles than the latter. Since the potential number of false positives incurred in this case is $> 1/3$ (Table 1), it is probably wise to take this putative association with a pinch of salt.

The map of relationships presented in Figure 19 provides an overview of the findings of our association study and includes several observations gained at the verification stage. In this diagram, positive associations are symbolized by green edges connecting the variables (phenotypes or genotypes) and negative associations by red edges. 'Weak edges', indicated by green or red dotted lines, represent associations thought to hold in a wide sense: for instance, *CETP Ile405Val* seems to be negatively associated with *most* of the five CETP polymorphisms listed in the box immediately below it, but in our study we have not actually declared it to be associated with *CETP (-1337)C>T*. In contrast, *CETP MspI* is linked to the box of those same five CETP polymorphisms by a 'strong edge' because it has been declared associated with *all* of them. Variables placed within green or red boxes indicate variables declared to be associated; for example, the five CETP polymorphisms we have been referring to are all thought to be positively associated with each other. ('Weak boxes', outlined by dotted rather than full lines, may be used in a wide sense for the same purpose, but we do not require them in the present diagram.) In contrast, *CETP Ile405Val* and *CETP (-972)G>A*, not having been declared associated, are placed within a black box. Age and BMI do not appear in the diagram, though as we have seen they affect the associations between the polymorphisms and other phenotypes. Other phenotypes such as MOD (obviously related to MSD) and activity of CETP (related with amount of CETP) also do not appear. Finally, some of the edges are supplemented with legends containing *post hoc* information, such as a question mark reminding us of the caveat about the putative association between family history and *PAI-1 (4G5G)*, or an indication about the age and BMI subgroups in which the corresponding associations are thought to be particularly strong. As examples of the latter, we mention the two CETP polymorphisms in the top left corner, which are thought to be associated with the amount of CETP in younger patients (no real evidence having been

DOI: 10.2202/1544-6115.1420

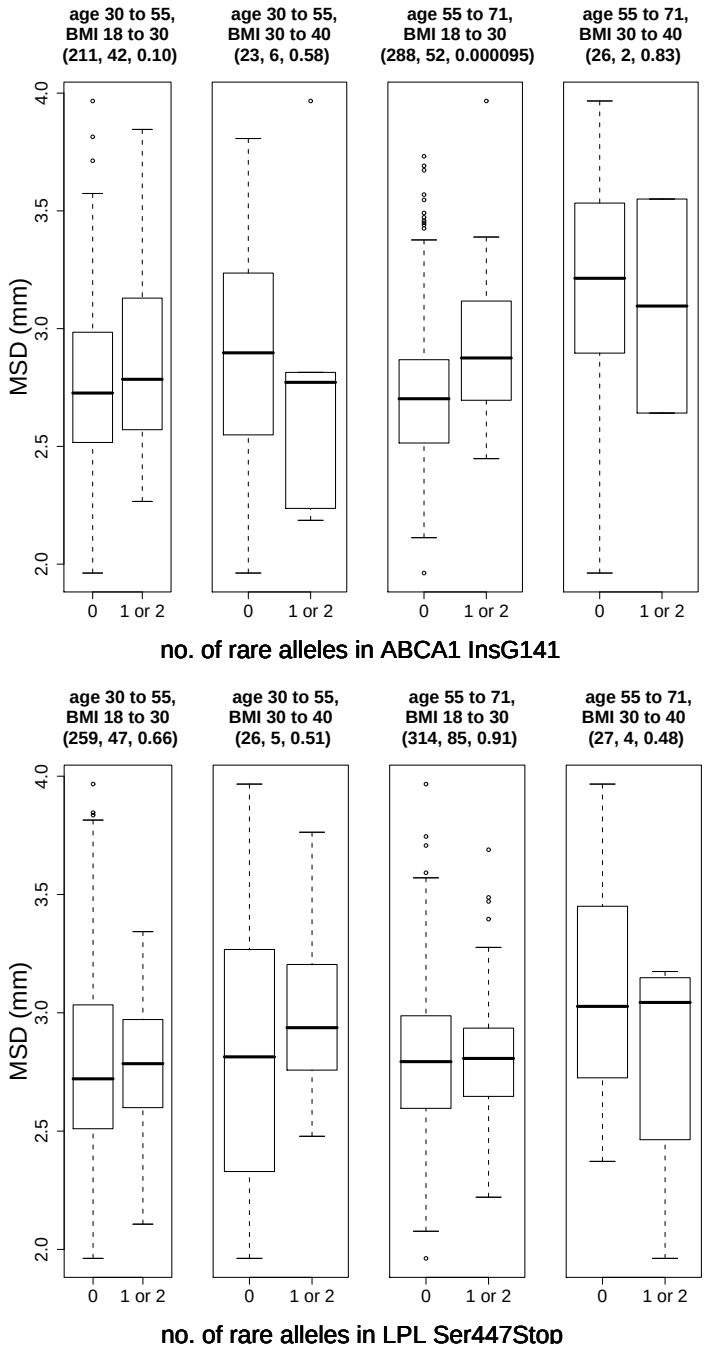

Figure 16: Boxplots of the distribution of MSD by number of rare alleles in the *ABCA1 InsG141* and *LPL Ser447Stop* polymorphisms stratified by age and BMI.

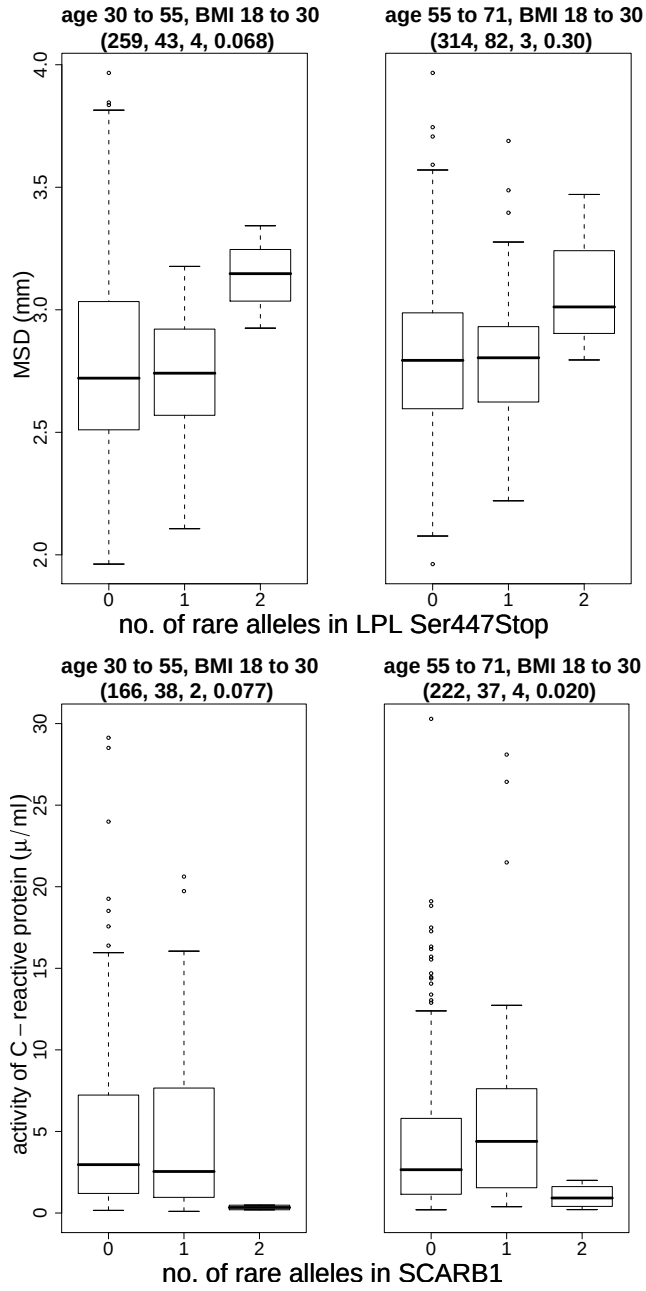

Figure 17: Boxplots of the distributions of MSD and C-reactive protein activity by number of rare alleles in the *LPL Ser447Stop* and *SCARB1* polymorphisms, respectively, in the bigger strata of BMI and age. (For the sake of visualization, some outliers have been omitted in the plots of C-reactive protein activity.)

DOI: 10.2202/1544-6115.1420

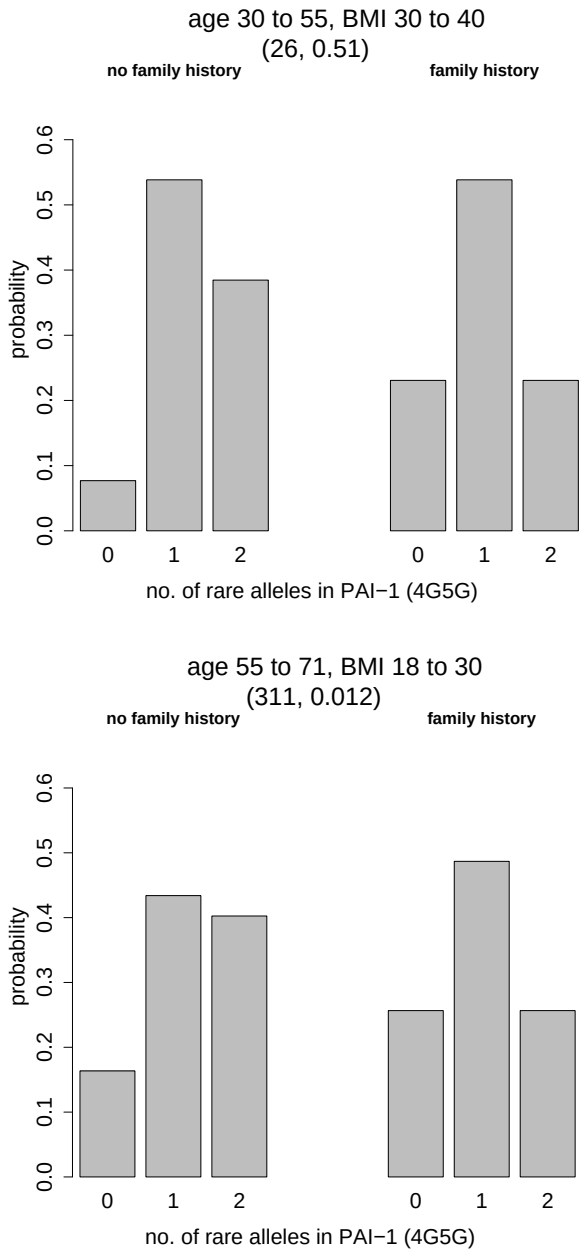

Figure 18: Histograms of the distribution of the number of rare alleles in the *PAI-1 (4G5G)* polymorphism stratified by age and BMI.

found in older patients), and the five CETP polymorphisms just below them, which seem to be associated with the phenotype especially (but not only) in young patients. The question and exclamation marks accompanying the links between polymorphisms of different genes (e.g. *CETP MspI* and the two ABCA1 polymorphisms) reflect the fact that besides being among the weakest in terms of p-values these associations are also the only unexpected ones from a biological point of view, which makes them likely candidates for false positives (though biologically there is no *a priori* reason to rule them out).

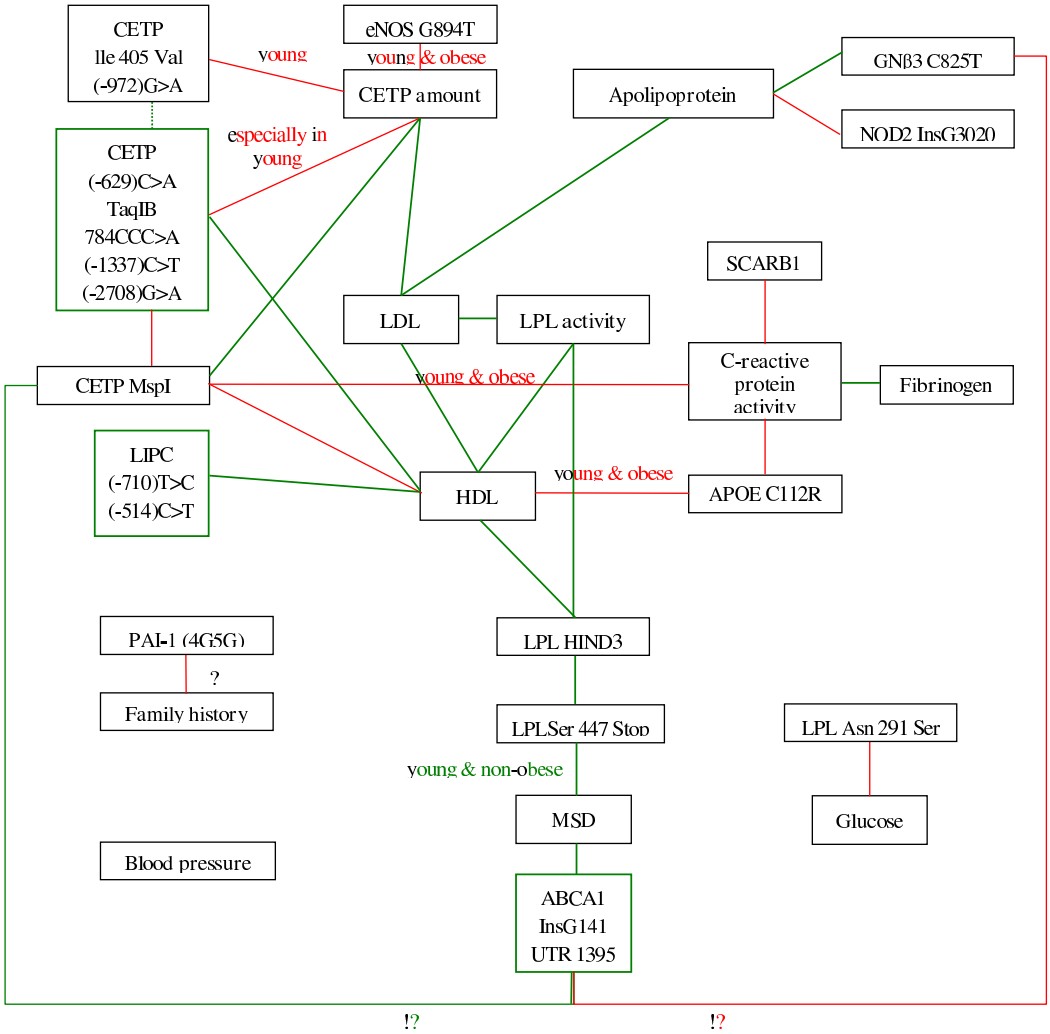

Figure 19: Map of relationships summarizing the main findings of the present association study: positive associations are indicated in green, negative associations in red.

DOI: 10.2202/1544-6115.1420

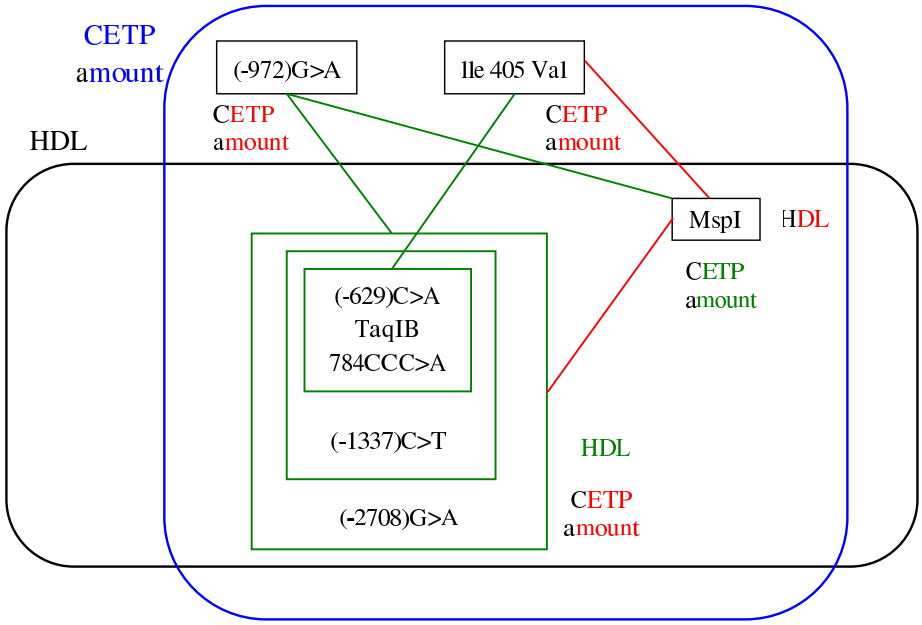

Figure 20: Map of relationships summarizing the associations between the CETP polymorphisms. The outlines in black and blue identify the subsets of polymorphisms putatively associated with HDL and amount of CETP.

For simplicity, the relationships between the CETP genes have not been fully represented in Figure 19, but they can be further visualized in Figure 20. There is a core of five polymorphisms positively associated with each other, positively associated with HDL, and negatively associated with amount of CETP. These five polymorphisms are in turn all negatively associated with *CETP MspI*, whose associations with HDL and CETP amount are, accordingly, negative and positive. This suggests that the six polymorphisms in the intersection of the black and blue outlines of Figure 20 function essentially as a unit regarding the two phenotypes. The remaining three polymorphisms probably deserve to be considered separately—for instance, *CETP Ile405Val* seems to be negatively associated with *CETP MspI* and positively associated with some of the 'core' CETP polymorphisms, and *CETP (-972)G>A* seems to be positively associated with *both CETP MspI* and the core five—and could perhaps help explaining why the effect of the CETP polymorphisms on CETP amount is greater in younger patients while their effect on HDL is, if anything, more visible in older patients.

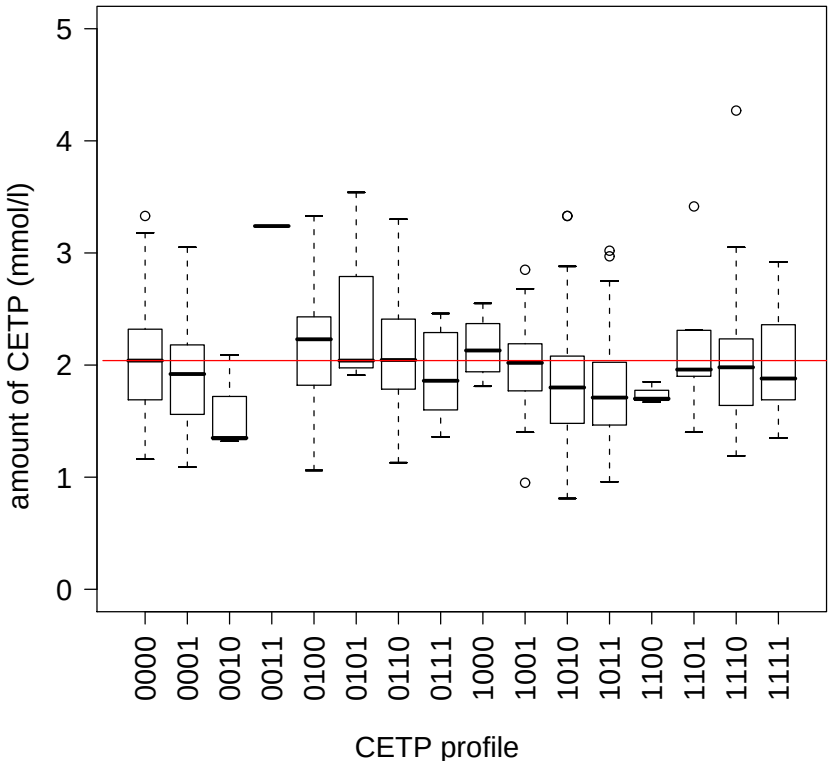

Figure 21: Boxplots of the distribution of CETP amount by 'CETP profile'.

Figure 20 prompts the question of whether anything can be learned from looking at the *joint effect* of polymorphisms on a phenotype. At this '*post post hoc*' stage there is probably very little new, *solid* information to be gained, and saying something substantial about how genetic interactions influence a given phenotype will certainly require data based on new, more refined studies. Nevertheless, for illustration purposes we shall finish our analysis by trying to extract something about the behaviour of amount of CETP as a function of a 'joint CETP profile'.

Taking into account what our diagrams say about the CETP polymorphisms and their associations with CETP amount, it seems sensible to create a CETP profile based on the eight polymorphisms contained in the blue outline of Figure 20 and to reduce the information on the core five polymorphisms *CETP (-629)G>A*, *TaqIB*, *784CCC>A*, *(-1337)C>T* and *(-2708)G>A* to a single variable. We shall therefore say that there is a 'mutation' in the core five polymorphisms if there is at least one rare allele in one of them, and to further simplify matters that there is a 'mutation' in each of the other

DOI: 10.2202/1544-6115.1420

three polymorphisms if there is at least one rare allele. Thus, indicating the existence of no rare alleles by 0 and the existence of rare alleles by 1, a profile will be represented by a string like 0100, where the first digit corresponds to the core five and the remaining to *CETP MspI*, *(-972)G>A* and *IIe405Val*, in this order. Figure 21 shows boxplots of the distribution of CETP amount as a function of the CETP profile thus created. The horizontal line is at the height of the median of the group with no mutations and may serve as a reference with which to compare the other groups. Going from left to right, we see once more that rare alleles in *(-972)G>A* and *IIe405Val* decrease the amount of CETP; rare alleles in MspI increase it; the increase in the phenotype 'due to' rare alleles in MspI is tempered by rare alleles in *(-972)G>A* and *IIe405Val*; two or three rare alleles occurring simultaneously in the core, *(-972)G>A* and *IIe405Val* polymorphisms seem to cause a bigger decrease in the amount of CETP, but even in such cases a rare allele in *CETP MspI* brings the levels of the phenotype close to 'normal'.

3.5 Discussion of the findings

Most of the associations between the 21 polymorphisms listed in Table 1 and various phenotypes have been studied by many authors and in particular by the authors of the Regress study. Associations between the CETP gene variants, CETP activity and HDL cholesterol have been observed in a previous analysis of the Regress data (Klerkx et al. (2003)) and in many others as well (e.g. Drayna and Lawn (1987), Funke et al. (1994), Bruce et al. (1998), Yamashita et al. (2000)). Consequently, several CETP-inhibitors have been developed for treatment of patients with low HDL cholesterol; recently one of these inhibitors was shown to be very effective in increasing HDL cholesterol but was also associated with increased mortality (Barter et al. (2007)). In a further analysis of the Regress data, Boekholdt et al. (2006) considered the association of HDL cholesterol with genetic variation in seven genes involved in reverse cholesterol transport. In addition to associations with the CETP gene, they found weak associations with markers in the lipoprotein lipase (LPL) and the scavenger receptor (SCARB1) genes but not with the hepatic lipase gene (LIPC), which we have singled out here. Indeed, when testing the association between HDL and *LIPC (-514)C>T* by means of analysis of variance they considered only the 546 individuals with no missing values in any of a given set of 14 polymorphisms rather than the group of 799 individuals with no missing values in HDL and *LIPC (-514)C>T*, and consequently obtained a p-value of 0.10 in place of the 'legitimate' p-value of 0.006, which is close to the one we got with the Scholz-Stephens test. This exemplifies

well one disadvantage of the regression approach to the detection of associations, for it was with the purpose of fitting a regression model to HDL as a function of the number of rare alleles that Boekholdt et al. (2006) discarded 32% of useful (as far as HDL and *LIPC (-514)C>T* are concerned) data. The APOE gene has been associated with LDL cholesterol many times but also with HDL cholesterol, either directly or via the association between HDL and LDL cholesterol (Burman et al. (2008)). Surprisingly, LDL cholesterol levels were not found to be associated with any genetic marker in the present analysis, while their heritability is thought to be high (O'Connell et al. (2005)) and several genes are known to influence LDL cholesterol's expression. Our study, however, includes only male patients with total cholesterol levels between 4 and 8 mmol/L, and we suspect that this selection severely limits LDL cholesterol variation and therefore the power of finding associations with genetic markers. Also surprising is the lack of association between systolic blood pressure and genetic variants, while many genes (for instance in the RAS system; see Dahlöf (2008)) have been described as influencing blood pressure. Apoliproprotein(a) levels are known to be very highly regulated by genetic variants in the LPA gene on locus 6q27; heritability of Apo(a) is thought to be high and mostly explained by this locus (Hasstedt et al. (1983)). Recently, several genome wide linkage studies (Lòpez et al. (2008)) identified several loci also influencing apo(a) among which are loci involved in the tissue factor pathway; GNβ3 and NOD2, however, have not been associated with Apo(a) before. Associations between inflammatory and thrombotic markers such as C-reactive protein and fibrinogen and serum lipids such as HDL and LDL cholesterol have been observed by many authors (Abou-Raya et al. (2007), Alber et al. (2008), and references therein). Associations between extent/severity of coronary artery disease and the markers listed in Table 1 have been studied by many authors. Hundreds of association studies with candidate genes (among which are APOE, LIPC, PPARG, TNF, PECAM, ABCA1, LPL) have been reported, but few have been convincingly confirmed. Possible explanations for this include publication bias, ethnic admixture, and small sample sizes. Genome-wide linkage studies have mapped several loci that may affect susceptibility to coronary artery disease/myocardial infarction (Watkins and Farrall (2006)), though only in two studies has the likely gene been identified (Helgadottir et al. (2004, 2006)). Association studies have identified several plausible genetic variants affecting lipids, thrombosis, inflammation or vascular biology, but for the most part the evidence is not yet conclusive (Watkins and Farrall (2006)). Recently, the Wellcome trust case-control consortium (and others) performed genome wide association analyses that did not find strong evidence for any of the candidate genes (Wellcome Trust Case Control Consortium (2007)).

DOI: 10.2202/1544-6115.1420

4 Discussion

The fact that we have singled out most of the polymorphisms that crop up in the literature in connection with heart disease, and even a couple of new ones, in a single scrutiny of the REGRESS dataset suggests that our approach, despite its emphasis on validity, can be powerful. Whether it is generally more or less powerful than other methods is a legitimate but futile question: for instance, comparing our method with some sort of parametric version of it will yield a whole range of different conclusions concerning detection power and control of the false discovery rate; this is obvious from the point of view of simulation, as one can vary the data generating model from the situation where the assumptions required by the parametric methods are fullfilled in all perfection to situations where they are grossly violated, and it is equally obvious from the point of view of real data, as different datasets will be compatible with the assumptions to a different degree and the truth about a dataset cannot really be ascertained.[9] The reasons why ours may be preferable to the usual approaches to association studies have been given in Section 2; they have mainly to do with transparency and with the need for clamping down on false positive results.

The verification and overview of the findings presented in Subsection 3.4, which culminates in the drawing up of the map of relationships, is an important part of our association study. Ideally, it should help biomedical researchers in elucidating the mechanisms of a disease—or at least in reformulating questions and setting up more specific studies with a view towards elucidating the mechanisms of a disease. The impression we have from reading the literature is that association studies—and we mean those addressed to biomedical audiences in the first place—tend to stop where statistical methods stop, namely at the point where a list of significant results is presented and observations are made about some of the significant results having been called significant by other studies, very much like in our discussion of the findings in Subsection 3.5. Compounded by Rees's (2002) opinion that "it is the geneticists and biochemists who need to learn some medicine", this would suggest that a lot more time and effort should be invested by both generalists and specialists in pure biomedical thinking going beyond the list of significant results. Our hope is that maps of relationships based on robust methods like

[9]It may be worth elaborating: If method A yields more 'discoveries' than method B at the same *conservative* false discovery rate on a given real dataset, this does not necessarily mean that A is better than B on that dataset, because on the one hand it is difficult to tell which discoveries are genuine and on the other hand method A may entail an actual false discovery well above that of B, which would explain the apparent difference in power.

that of Figure 19 will provide not only reliable and digestible summaries of statistical results but also ground for thorough, properly biomedical investigations.

It must be recognized, however, that the feasibility of a map of relationships depends very much on the problem and dataset at hand. For example, it is doubtful whether the results of a study involving thousands or tens of thousands of 'genes' could be summarized by a couple of diagrams without some sort of downscaling procedure. In principle, the first three steps of our approach can cope with any number of explanatory variables, but the fourth step ('verification and overview') would definitely need to be expanded to deal with cases where the number of associations declared significant is in the order of hundreds or thousands.

Appendix: List of gene polymorphisms

No.	Name in SAS file	Gene name	Abbreviation	Polymorphism
1	AA1_83	Apolipoprotein A1	APOA1	C83T
2	AA1_M75		APOA1	(-75)G>A
3	AA4_347	Apolipoprotein A4	APOA4	Thr347Ser
4	AA4_360		APOA4	Gln360His
5	AA5_pol1	Apolipoprotein A5	APOA5	(-1131)T>C
6	AA5_S19W		APOA5	Ser19Trp
7	AAT_213	Protein inhibitor	PI	V213A
8	AAT_TAQI		PI	G1237A?
9	AB_3500	Apolipoprotein B	APOB	Arg3500Gln
10	AB_71		APOB	Thr71Ile
11	ABC_1051	ATP-binding cassette transport protein A1	ABCA1	G1051A
12	ABC_1252		ABCA1	(-1252)G>A
13	ABC_1320		ABCA1	?
14	ABC_1395		ABCA1	(-1395)C>T
15	ABC_141		ABCA1	?
16	ABC_1591		ABCA1	T1591C
17	ABC_1699		ABCA1	?
18	ABC_17		ABCA1	C17G
19	ABC_191		ABCA1	(-191)G>C
20	ABC_217		ABCA1	?
21	ABC_2706		ABCA1	G2706A
22	ABC_2715		ABCA1	A2715C
23	ABC_2723		ABCA1	G2723C
24	ABC_2868		ABCA1	G2868A
25	ABC_3044		ABCA1	A3044G
26	ABC_3911		ABCA1	G3911C
27	ABC_5155		ABCA1	G5155A
28	ABC_518		ABCA1	?
29	ABC_5587		ABCA1	C5587G
30	ABC_6844		ABCA1	C6844T
31	ABC_69		ABCA1	C69T
32	ABC_877		ABCA1	Ala877Val
33	ABC_M477		ABCA1	(-477)C>T
34	AC3_1100	Apolipoprotein C3	APOC3	C1100T
35	AC3_3175		APOC3	C3175G, SstI
36	AC3_3206		APOC3	T3206G
37	AC3_M455		APOC3	(-455)T>C
38	AC3_M482		APOC3	C(-482)T
39	AC3_M641		APOC3	(-641)C>A
40	ACE	Angiotensin I converting enzyme	ACE	Ins287 bp Alu repeat sequence
41	AD1_460	Adducin 1	ADD1	G460W
42	AE_112	Apolipoprotein E	APOE	Cys112Arg
43	AE_158		APOE	Arg158Cys
44	ANG_174	Angiotensinogen	AGT	T174M
45	ANG 235		AGT	M235T

DOI: 10.2202/1544-6115.1420

No.	Name in SAS file	Gene name	Abbreviation	Polymorphism
46	ANP_2238	Natriuretic Peptide Precursor A	NPPA	T2238C
47	ANP_7		NPPA	V7M
48	AR2_16	Beta-2-adrenergic receptor	ADRB2	R16G
49	AR2_27		ADRB2	Q27E
50	AR3_64	Beta-2-adrenergic receptor	ADRB3	W64R
51	ATI_1166	Angiotensin II receptor type 1 (ATII)	AGTR1	A1166C
52	BN_648	Unknown HDL modifier gene	BRUNHAM	A648G
53	BN_726		BRUNHAM	C726A
54	C_2	Cholesterol ester transport protein	CETP	G1A
55	C_405		CETP	Ile405Val
56	C_442		CETP	Asp442Gly
57	C_784		CETP	CCC784A
58	C_ECONI		CETP	ECONI
59	C_M1337		CETP	(-1337)C>T
60	C_M1614		CETP	A6-1614A7
61	C_M1632		CETP	A6-1632A5
62	C_M1653		CETP	(-1653)T>C
63	C_M1673		CETP	(-1673)T>C
64	C_M1932		CETP	(-1932)T>C
65	C_M2708		CETP	(-2708)G>A
66	C_M629		CETP	(-629)C>A
67	C_M630		CETP	(-630)C>A
68	C_M827		CETP	(-827)C>T
69	C_M875		CETP	(-875)C>T
70	C_M972		CETP	(-972)G>A
71	C_MSPI		CETP	MspI
72	C_TAQIA		CETP	TAQIA
73	C-TAQIB		CETP	TAQIB
74	CBS_833	Cystathionine synthase	CBS	T833C (I278T)
75	CBS_INS		CBS	844ins68
76	CD_M159	CD14 lipopolysaccharides receptor	CD14	(-159)T>C
77	Cyp7	Cholesterol 7-alpha-hydroxylase	CYP7A1	A278C
78	ENO_894	Endothelial Nitric Oxide Synthase	eNOS	G894T
79	ENO_M690		eNOS	(-690)C>T
80	ENO_M786		eNOS	(-786)T>C
81	ENO_M922		eNOS	(-922)A>G ((-948)A>G?)
82	ENO_VNTR		eNOS	VNTR
83	F2_20210	Blood coagulation factor II (prothrombin)	F2	G20210A
84	F5_506	Blood coagulation factor V	F5	G1691A / Q506R
85	F7_10976	Blood coagulation factor VII	F7	G10976A
86	F7_10B		F7	10-BP INS, NT-323?
87	FGA_TAQI	Fibrinogen a	FGA	TAQI
88	FGB_M455	Fibrinogen	FGB	(-455)G>A
89	GN3_825	Guanine Nucleotide-binding protein, beta-3	GNB3	C825T
90	GP3_1565	Glycoprotein IIIa (Human platelet (allo)antigen, HPA)	GP IIIa	T1565C
91	GPI_807	Glycoprotein Ia	GP Ia	C807T
92	GPI_873		GP Ia	G873A
93	HL_M514	Hepatic lipase	LIPC	(-514)C>T
94	HL_M710		LIPC	(-710)T>C
95	HL_RFL19		LIPC	RFLP1?
96	HL_RFL20		LIPC	RFPL2?
97	IC_214	Intercellular adhesion molecule 1	ICAM1	?
98	LC_208	Lecithin-cholesterol acyltransferase	LCAT	Ser208Thr
99	LD_1170	LDL receptor gene	LDLR	?
100	LD_1959		LDLR	C1959T
101	LD_NCOI		LDLR	NCOI
102	LD_TAQI		LDLR	TAQI
103	LP_291	Lipoprotein lipase	LPL	Asn291Ser
104	LP_447		LPL	Ser447Stop
105	LP_9		LPL	Asp9Asn
106	LP_HIND3		LPL	HIND3
107	LP_M93		LPL	(-93)T>G
108	LP_PVUI		LPL	PVUI
109	LPA_121	Lipoprotein A, Lp(a)	LPA	G121A
110	LPA_93		LPA	C93T
111	LTA_26	Lymphotoxin-Alpha	LTA	Thr26Asn
112	MP_COMB	?	?	?
113	MP3_STRO	Matrix metalloproteinase 3	MMP3	?
114	MP9_MMP9	Matrix metalloproteinase 9 (Gelatinase B)	MMP9	?
115	MT_677	Methylenetetrahydrofolate reductase	MTHFR	C677T
116	MTP_493	Microsomal triglyceride transfer protein	MTP	(-493)G>T
117	NOD_3020	Nucleotide-binding oligomerization domain protein 2	NOD2	3020InsC
118	NOD_908		NOD2	G908R
119	NOD_Unkn		NOD2	?
120	PA_11053	Plasminogen Activator Inhibitor 1	PAI1	G11053T
121	PA_4g5g		PAI1	4G5G
122	PEC_125	Platelet/endothelial cell adhesion molecule (CD31 antigen)	PECAM1	Leu125Val

123	PEC_53		PECAM1	G53A
124	PEC_670		PECAM1	R670G
125	PON_192	Paraoxonase I	PON1	Gln192Arg
126	PON_55		PON1	Leu55Met
127	PON2_311	Paraoxonase II	PON2	Ser311Cys
128	PPA_162	Peroxisome proliferative activated receptor α	PPARa	Leu162Val
129	PPA_3		PPARa	G2176A
130	PPG_12	Peroxisome proliferative activated receptor gamma	PPARg	Pro12Ala
131	SCA_493	SODIUM CHANNEL, NONVOLTAGE--GATED 1, ALPHA SUBUNIT	SCNN1A SCNN1A	W493R
132	SCA_663		SCNN1A	?
133	SE_128	Selectin E	SELE	S128R
134	SE_554		SELE	Leu554Phe
135	SOD_213	Superoxide Dismutase (extracellular)	SOD3	Arg213Gly
136	SRB1_E1	Scavenger Receptor class B type 1	SRB1	Gly2Ser
137	SRB1_E8		SRB1	C1050T
138	SRB1_I5		SRB1	Intron 5 C/T
139	TL4_299	Toll-like receptor type 4	TLR4	D299G
140	TL4_399		TLR4	T399I
141	TNF_M238	Tumor necrosis factor alpha	TNFa	(-238)G>A
142	TNF_M244		TNFa	?
143	TNF_M308		TNFa	(-308)G>A
144	TNF_M376		TNFa	(-376)G>A

References

Abou-Raya, S., Abou-Raya, A., Naim, A., and Abuelkheir, H. (2007). Chronic inflammatory autoimmune disorders and atherosclerosis. *Ann. N.Y. Acad. Sci.*, *1107*(1).

Alber, H., Wanitschek, M., de Waha, S., Ladurner, A., Suessenbacher, A., Dörler, J., Dichtl, W., Frick, M., Ulmer, H., Pachinger, O., and Weidinger, F. (2008). High-density lipoprotein cholesterol, C-reactive protein, and prevalence and severity of coronary artery disease in 5641 consecutive patients undergoing coronary angiography. *Eur. J. Clin. Invest.*, *38*(6), 372-80.

Barter, P., Caulfield, M., Eriksson, M., Grundy, S., Kastelein, J., Komajda, M., Lopez-Sendon, J., Mosca, L., Tardif, J., Waters, D., Shear, C., Revkin, J., Buhr, K., Fisher, M., Tall, A., Brewer, B., and ILLUMINATE Investigators. (2007). Effects of torcetrapib in patients at high risk for coronary events. *N. Engl. J. Med.*, *357*(21), 2109-22.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*(57), 289-300.

Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, *25*.

DOI: 10.2202/1544-6115.1420

Benjamini, Y., Krieger, A., and Yekutieli, D. (2006). Adaptive Linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491-507.

Boekholdt, S., Souverein, O., Tanck, M., Hovingh, G., Kuivenhoven, J., Peters, R., Jansen, H., Schiffers, P., van der Wall, E., Doevendans, P., Reitsma, P., Zwinderman, A., Kastelein, J., and Jukema, J. (2006). Common variants of multiple genes that control reverse cholesterol transport together explain only a minor part of the variation of HDL cholesterol levels. *Clin. Genet.*, *69*(3), 263-70.

Bruce, C., Chouinard, R., and Tall, A. (1998). Plasma lipid transfer proteins, high-density lipoproteins, and reverse cholesterol transport. *A. Rev. Nutr.*, *18*, 297330.

Burman, D., Mente, A., Hegele, R., Islam, S., Yusuf, S., and Anand, S. (2008). Relationship of the ApoE polymorphism to plasma lipid traits among south asians, chinese, and europeans living in canada. *Atherosclerosis*. ([Epub ahead of print])

Dahlöf, B. (2008). Management of cardiovascular risk with RAS inhibitor/CCB combination therapy. *J. Hum. Hypertens.* (In press)

Drayna, D., and Lawn, R. (1987). Multiple RFLPs at the human cholesteryl ester transfer protein (CETP) locus. *Nucl. Acids Res.*, *15*, 4698.

Dudoit, S., Gilbert, H., and van der Laan, M. (2008). Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: focus on the false discovery rate and simulation study. *Biometrical Journal*, *50*(5), 716-44.

Ferreira, J., and Zwinderman, A. (2006). Approximate power and sample size calculations with the Benjamini-Hochberg method. *The International Journal of Biostatistics*, *Vol. 2, Issue 1, Article 8*. (Available online at http://www.bepress.com/ijb/vol2/iss1/8)

Funke, H., Wiebusch, H., Fuer, L., Muntoni, S., Schulte, H., and Assmann, G. (1994). Identification of mutations in the cholesteryl ester transfer protein in Europeans with elevated high density lipoprotein cholesterol. *Circulation*, *90*.

Goodman, S., and Greenland, S. (2007). Assessing the unreliability of the medical literature: a response to "Why most published research findings are false". (Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 135 available at http://www.bepress.com/jhubiostat/paper135)

Hasstedt, S., Wilson, D., Edwards, C., Cannon, W., Carmelli, D., and Williams, R. (1983). The genetics of quantitative plasma Lp(a): analysis of a large pedigree. *Am. J. Med. Genet.*, *16*, 179-188.

Helgadottir, A. et al. (2004). The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat. Genet.*, *36*, 233-239.

Helgadottir, A. et al. (2006). A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat. Genet.*, *38*, 68-74.

Helland, I. (1995). Simple counterexamples against the conditionality principle. *The American Statistician*, *49*(4), 351-356.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), 696-701.

Jukema, J., Bruschke, A., Boven, A. van, Reiber, J., Egbert, T., Zwinderman, A., Jansen, H., Boerma, G., Rappard, F. v., and Li, K. (1995). Effects of lipid lowering by pravastatin on progression and regression of coronary heart disease in symptomatic men with normal to moderately elevated serum cholesterol levels: the regression growth evaluation statin study (REGRESS). *Circulation*, *91*, 2528-2540.

Kathiresan, S., Manning, A., Demissie, S., D'Agostino, R., Surti, A., Candace, G., Gianniny, L., Burtt, N., Melander, O., Orho-Melander, M., Arnett, D., Peloso, G., Ordovas, J., and Cupples, L. (2007). A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Medical Genetics*. (Available online at http://www.biomedcentral.com/I 47 I-2350/8/SI/SI 7)

Kim, K., and Wiel, M. van de. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, *9*(114).

DOI: 10.2202/1544-6115.1420

Klerkx, A., Tanck, M., Kastelein, J., Molhuizen, H., Jukema, J., Zwinderman, A., and Kuivenhoven, J. (2003). Haplotype analysis of the CETP gene: not TaqIB, but the closely linked $-629C \rightarrow A$ polymorphism and a novel promoter variant are independently associated with CETP concentration. *Hum. Mol. Genet.*, *12*(2), 111-23.

Kullo, I., and Ding, K. (2007). Mechanisms of disease: the genetic basis of coronary heart disease. *Nature Clinical Practice: Cardiovascular Medicine*, *4*(10), 558-569.

Langaas, M., Lindqvist, B., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. Roy. Stat. Soc. Ser. B*, *8*(4), 555-572.

Lòpez, S., Buil, A., Ordoez, J., Souto, J., Almasy, L., Lathrop, M., Blangero, J., Blanco-Vaca, F., Fontcuberta, J., and Soria, J. (2008). Genome-wide linkage analysis for identifying quantitative trait loci involved in the regulation of lipoprotein a (lpa) levels. *Eur. J. Hum. Genet.* ([Epub ahead of print])

Maat, M., Kastelein, J., Jukema, J., Zwinderman, A., Jansen, H., Groenemeier, B., Bruschke, A., and Kluft, C. (1998). -455g/a polymorphism of the β-fibrinogen gene is associated with the progression of coronary atherosclerosis in symptomatic men. *Artherioscler. Thromb. Vasc. Biol.*, *18*, 265-271.

O'Connell, D., Heller, R., Roberts, D., Allen, J., Knapp, J., Steele, P., Silove, D., Vogler, G., and Rao, D. (2005). Twin study of genetic and environmental effects on lipid levels. *Genetic Epidemiology*, *5*(5), 323-341.

Rees, J. (2002). Complex disease and the new clinical sciences. *Science*, *296*, 698-701.

Romano, J., Shaikh, A., and Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, *17*(3), 417-442.

Rothman, K. (1986). *Modern Epidemiology.* Little, Brown and Company.

Rothman, K. (2002). *Epidemiology: an introduction.* New York: Oxford University Press.

Scholz, F., and Stephens, M. (1987). K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, *82*(399), 918-924.

Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Stat. Soc. Ser. B*(66), 187-205.

Storey, J., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, *100*(16), 9440-9445.

Watkins, H., and Farrall, M. (2006). Genetic susceptibility to coronary artery disease: from promise to progress. *Nat. Rev. Genet.*, *7*, 163-173.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661-78.

Willett, W. (2002). Balancing life-style and genomics research for disease prevention. *Science*, *296*, 695-698.

Winkelmann, B., Hager, J., Kraus, W., Merlini, P., Keavney, B., Grant, P., Muhlestein, J., and Granger, C. (2000). Genetics of coronary heart disease: Current knowledge and research principles. *American Heart Journal*, *1*(40), S11-S26.

Yamashita, S., Hirano, K., Sakai, N., and Matsuzawa, Y. (2000). Molecular biology and pathophysiological aspects of plasma cholesteryl ester transfer proteins. *Biochim. Biophys. Acta*, *1529*, 257275.

Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Society*, *103*(481), 309-16.

Yekutieli, D., Reiner-Benaim, A., Benjamini, Y., Elmer, G., Kafkafi, N., Letwin, N., and Lee, N. (2006). Approaches to multiplicity issues in complex research in microarray analysis. *Statistica Neerlandica*, *60*(4), 411-437.

DOI: 10.2202/1544-6115.1420