Original Paper

Human Heredity

Hum Hered 2010;70:34–41
DOI: 10.1159/000293722

Received: December 7, 2009
Accepted after revision: March 1, 2010
Published online: April 23, 2010

Genome-Wide Conditional Search for Epistatic Disease-Predisposing Variants in Human Association Studies

Gao Wang[a, c] Yaning Yang[b] Jurg Ott[a]

[a]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, and [b]Department of Statistics and Finance, University of Science and Technology of China, Hefei, China; [c]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Tex., USA

Key Words
Genome-wide association analysis · Conditional search · Epistatic disease-predisposing variants

Abstract
Genome-wide search for new disease variants, based on well-established variants, has a long history in linkage analysis but is less well-known in genetic case-control association studies. We developed a simple yet highly efficient conditional search method that can find new variants, which are associated with a disease only through epistatic interaction with another variant and do not necessarily have a direct association effect. Our approach is analogous to partitioning of χ^2 in a hierarchical design, which is a well-established statistical technique. Applied to previously published data on age-related macular degeneration, our method found two single-nucleotide polymorphisms with genome-wide significant epistatic interaction that could not be found based only on direct main effects.

Introduction

Even though association studies are traditionally based on direct (main) effects of individual single-nucleotide polymorphisms (SNPs), they have been successful in a number of applications [1, 2]. However, variants with low or absent direct main effects cannot generally be discovered by these methods. Here we demonstrate simple yet powerful methods for detecting variants through their epistatic associations with strong loci, even though these variants have low main effects themselves. That is, we aimed to detect disease-causing variants that cannot be found on the basis of their own association with a disease and can only be 'seen' via their association with stronger loci. Thus, we propose a conditional search in case-control association studies by initially searching for associated variants based on their direct (main) effects (stage 1), and then by searching the genome again but conditional on genotypes of the previously discovered SNPs (stage 2). Our aim is to combine stages 1 and 2 in a permutation testing framework to obtain a significance level corrected for testing multiple SNPs and working with two stages.

In the context of genetic linkage analysis, several strategies have been distinguished to localize disease suscep-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/hhe

Jurg Ott
Beijing Institute of Genomics, Chinese Academy of Sciences
7 Bei Tu Cheng West Road
Beijing 100029 (China)
Tel./Fax +86 10 8299 5385, E-Mail ottjurg@yahoo.com

tibility variants, for example, single-locus search, simultaneous search, and conditional search [3], where simultaneous search methods have previously been proposed [4], and likelihoods at one locus conditional on the phenotype at another locus have been proposed long ago to allow for epistatic interactions between loci [5]. These principles were first applied in human genome-wide linkage searches by subdividing affected sibpairs into two groups depending on the identity-by-descent status at HLA [6] or depending on HLA phenotypes [7], and have since been used widely in linkage mapping [8–10]. For family data, one can condition either on genotypes/phenotypes or on IBD statuses at another locus. These two strategies are very different and conditioning on IBD statuses is unique to family data. Di and Thompson [11] discussed how to perform conditional tests based on IBD statuses for extended families.

In genome-wide case-control association studies (GWAS), to allow for the simultaneous effects of multiple susceptibility variants, two-stage approaches have been proposed with the aim to select good candidate variants in stage 1 and then carry out joint analysis in step 2 [12, 13]. Another approach consists of analysis in a single step, but the joint effects of multiple SNPs (at different genomic locations) is approximated by the sums of single-locus test statistics [14]. A more recent two-stage strategy compares genotype pattern (diplotype) frequencies in case and control individuals at a number of SNPs that have been picked in a GWAS for their individual main effects [15].

Many of these approaches focus on main effects of single SNPs. In complex traits, it is generally believed that multiple interacting variants contribute to disease susceptibility [16, 17]. Multilocus search methods have been proposed [18] and are powerful but, because of the daunting number of SNP combinations, they are not applicable to genome-wide searches. An exhaustive two-locus search for disease-associated variants has been proposed and shown to be feasible [13]. Recently, Bayesian approaches have been described for analyzing all SNPs in human [19–21] and animal [22, 23] association studies. As an alternative approach, we propose a conditional search, which will allow the detection of disease-associated SNPs with low or absent main effects as long as they act in concert with candidate genes or with variants that exert direct main effects on the occurrence of a disease, where these main effects need not be statistically significant. Our strategy may be called a two-stage or conditional search. We prefer the latter term as previously introduced, for example, in the situation that 'we have performed a preliminary analysis, for example, a single locus search, and have detected linkage to some loci and would like to continue analysis of the same data to find evidence for linkage to other loci' [3].

As will be seen below, our methods are applicable to both positive and negative interactions. The latter have been termed antagonistic interactions or negative regulation. In human data, few examples of this type of interaction have been documented [24–26]. Also, for example, modifier genes are known to increase or decrease the risk or speed up or delay the time of onset of AIDS [27, 28].

Stratification of Data

The idea underlying our approach is simple and has previously been applied in an ad hoc manner [6, 7]. Given a SNP (here called test SNP) with known or suspected disease association, we aimed to find a variant (here called target SNP) causing a disease through its association with the test SNP and its own possibly small main effect. We stratified the data using the three genotypes of the test SNP and carried out an association analysis for each of the resulting three subsets of the data. Genotypes at the test and target SNPs form a 3×3 contingency table with a total of 8 degrees of freedom (d.f.), 2 d.f. for each of the two main effects and 4 d.f. for the interaction between the two SNPs. Epistatic effects may be detected by a 4-d.f. test for interaction [29] but, as outlined below, here we propose a 6-d.f. test for interaction and main effect at the target SNP.

To see the motivation for our approach, consider a two-locus model previously proposed to show strong epistatic interaction in the absence of main effects at the two loci [30]. We modified this model in the following manner. The test locus has genotypes *AA*, *AB*, and *BB*, with allele frequency $p = P(B) = 0.50$, while the target locus has genotypes *CC*, *CD*, and *DD* with allele frequency $q = P(D) = 0.20$, where genotype frequencies are given by the Hardy-Weinberg law. Penetrances at the nine two-locus genotypes are shown in table 1 and are the same as in the original publication [30]. For given sample sizes of cases and controls, n = 20 each, we compute expected numbers of observations at the nine genotypes in case and control individuals. The main (marginal) effect of a locus is then given by the likelihood ratio (LR) χ^2 obtained from a 2×3 table of genotypes, where the two rows correspond to cases and controls and the three columns represent the three genotypes, while the body of the table contains expected numbers of observations. The resulting expected

numbers, based on the penetrances in table 1, lead to χ^2 (2 d.f.) values of 12.78 and 0 for the test and target loci, respectively. Thus, the target locus shows no main effect while the test locus main effect is rather strong, so that a 'regular' case-control association analysis has no chance of ever detecting the target locus on its own. However, when the data are subdivided into the three genotypes at the test SNP, the resulting 2 × 3 genotype tables each show strong conditional effects of the target locus. The expected χ^2 values (2 d.f. each) are 7.20, 10.69, and 15.20 for genotypes *AA*, *AB*, and *BB*, respectively. Thus, this simple device of stratifying the data by test locus genotypes has the potential to exhibit target locus effects that are hidden in the overall picture of things. For an in-depth discussion of purely epistatic inheritance models, see Culverhouse et al. [31].

Table 1. Penetrances adapted from a modified Frankel and Schork [30] model of two epistatically interacting trait variants with allele frequencies of P(*B*) = 0.50 and P(*D*) = 0.20

Test locus	Target locus			Marginal penetrance
	CC	*CD*	*DD*	
AA	0	0	1	0.04
AB	0	0.5	0	0.16
BB	1	0	0	0.64
Marginal penetrance	0.25	0.25	0.25	

Table 2. Numbers of genotypes at two SNPs in a genome-wide case-control study of age-related macular degeneration [38]

Test:	*AA*			*AB*			*BB*		
Target:	*CC*	*CD*	*DD*	*CC*	*CD*	*DD*	*CC*	*CD*	*DD*
Cases	1	1	4	19	31	8	12	9	11
Controls	3	3	0	14	14	8	30	53	1

Test SNP = SNP_A-1702501, target SNP = rs1957491.

Table 3. Partitioning of χ^2 for the data in table 2

Source	χ^2	d.f.
Test SNP main effect	25.2921	2
Target SNP main effect	12.8013	2
Interaction by subtraction	25.1310	4
Total	63.2244	8

Methods

Test Statistics for Data Stratification

Assume that in the data for a given test locus genotype (*AA*, *AB*, or *BB*), we test for association by χ^2 statistic in a 2 × 3 genotype table with the rows corresponding to cases and controls and the three columns representing the target SNP genotypes *CC*, *CD*, and *DD*. For an overall assessment of the association between target locus and disease, given the test locus genotypes, we aim to combine the three (independent) test statistics in a suitable manner. The two simplest statistics are the sum, T_{sum}, and the maximum, T_{max}, of the three χ^2. Under H_0, T_{sum} follows a χ^2 distribution with 6 d.f. The significance level of T_{max} is given by the p value, p_{min}, of the largest of the three χ^2. As three independent data sets are tested, the probability that one or more of them show a χ^2 exceeding T_{max} is given by $1 - (1 - p_{min})^3$ (Bonferroni-type correction for multiple testing).

To our knowledge, only one conditional search method, the overall conditional genotype (OGT) method [32, 33], has previously been proposed with the aim to find additional genes in the HLA region based on previously detected HLA variants. It applies the same principle of data stratification but, based on marginal genotype frequencies of controls at the target locus, it estimates marginal expected numbers of cases and compares them with observed marginal numbers of genotypes at the target locus. This comparison is then tested with a χ^2 statistic, here referred to as T_{OGT}. Its definition is fairly involved and has explicitly been given elsewhere [32].

Partitioning of χ^2

Consider a 2 × 9 contingency table, the two rows of which correspond to cases and controls, and the nine columns represent all pairs of genotypes at two SNPs. χ^2 (8 d.f.) for this table comprises the main association effects of the two SNPs and all interaction effects between the two SNPs. We view the layout of such a 2 × 9 table as a hierarchical design [34]. The data are first stratified by a factor with three levels, *AA*, *AB*, and *BB* (the three genotypes at the test locus), and then further subdivided within each stratum. In analogy to the probability formula $P(L_1, L_2) = P(L_1)\,P(L_2|L_1)$, where L_1 refers to the test locus and L_2 to the target locus genotypes, the total χ^2 for the 2 × 9 table may be partitioned into two components [34], one due to the (main) effects of the test SNP and the other due to main effects at the target locus and its interactions with the test locus. A specific example is shown in table 2 and referred to again in the Application section. This partitioning of χ^2 is a standard approach in hierarchical designs [34]. It is evident from its construction that the test statistic, T_{sum}, is identical with the sum of the two χ^2 due to 'Target SNP main effect' and 'Interaction' (table 3), which may be obtained as the difference between the total χ^2 for the 2 × 9 table and the main effect χ^2 from the marginal 2 × 3 table of the test SNP. Thus, data stratification in genetic analysis has a well-established statistical foundation.

At this point, we are ready to formulate the null hypothesis, H_0, for our test: absence of 'Target SNP main effect' and absence of 'Interaction'.

In case-control studies, tests for epistatic interaction between two SNPs may be formulated within the framework of logistic regression [29, 35]. Because of the asymptotic equivalence between χ^2 tests and logistic regression analysis, our T_{sum} statistic may also be obtained via logistic regression, which is discussed below.

Power Simulations

In order to see the usefulness of the test statistics considered here (T_{sum}, T_{max}, and T_{OGT} mentioned above) and to compare their power to detect a target SNP, we carried out computer simulations under controlled statistical conditions. Thus, we generated case and control data (n = 200 each, except where noted otherwise) for several common inheritance models and analyzed them with each of the three test statistics, choosing a threshold for significance such that power = 0.05 under the null hypothesis. As mentioned, our interest is in models with weak or absent main effects of the target variant. The different models are generally calibrated to predict a population prevalence of 5%, that is, they represent a fairly common genetic trait. The power simulations reported below refer to stage 2 analysis and do not take into account how test SNPs were obtained at stage 1. That is, genotypes are generated for two SNPs, the test and target SNP. In practice, of course, researchers will generally work with large numbers of SNPs (see Practical Application below), but here we want to compare relative power for different analysis methods rather than absolute power in a genome-wide setting.

Adapted from the modified Frankel and Schork [30] model demonstrated in table 1, power (y-axis) in figure 1 is shown as a function of $1 - q$ (x-axis), where q is the allele frequency P(D) at the target SNP. The main effect of the test SNP is zero at $q = 0.5$ and increases with increasing value of $1 - q$. In this model, T_{max} is most powerful, closely followed by T_{OGT}, while T_{sum} is much less powerful. This model is perhaps not very realistic, and we use it mainly to document that our method can easily find a disease susceptibility variant that has no main effect but exerts disease association through its strong epistatic interaction with the test locus.

Next, we considered a logistic regression model with increasing interaction effects and target SNP main effects, where the latter are at most 1/6 of those at the test SNP. The model equation reads

$$\log[\phi/(1-\phi)] = c_0 + c_1x_1 + c_2x_2 + s[c_1x_3 + c_2x_4]/6 + sc_1(x_1 + x_2)(x_3 + x_4)/2,$$

where ϕ is the conditional probability of being affected given model parameters. The two d.f. of the test SNP genotypes are modeled by two independent dummy variables, $x_1 = (-1, 0, +1)$ and $x_2 = (-1, +2, -1)$, with x_3 and x_4 being the corresponding quantities for the target locus [29, 36]. We arbitrarily defined $c_1 = 1.2$ and $c_2 = 0.2$, and adjusted c_0 to achieve a population prevalence of 5%. Given these parameter settings, we computed power as a function of s, with $s = 0$ representing the absence of interaction effects and target locus main effects. As seen in figure 2, the T_{sum} statistic is most powerful, closely followed by T_{max}, while T_{OGT} has considerably less power.

Finally, we adopted the two-locus multiplicative interaction model from figure 1 in Marchini et al. [13], with $\alpha = 0.1$, $\theta = 0.5$, and n = 500 cases and controls each, where α are baseline odds for a disease and θ represents a genotypic effect. We take the A locus to be our test SNP and compute power (y-axis in fig. 3) as a function of its allele frequency (x-axis), with P(A) = 0.05 as the baseline. The target locus (B locus [13]) with a fixed allele frequency of 0.5 has little main effect. Here again, T_{sum} has most power, while the other two statistics are clearly less powerful.

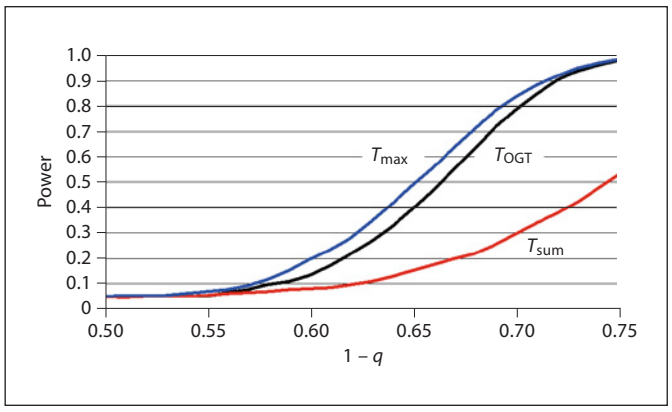


Fig. 1. Power (y-axis) as a function of $1 - q$ (q = allele frequency of target locus) of three conditional test statistics adapted from the modified Frankel and Schork [30] model shown in table 1.

We also investigated the power for a few other genetic models (not shown here) and generally found the T_{OGT} statistic to be the least powerful for our purpose of finding variants with low or absent main effects. The T_{sum} statistic often had somewhat more power than T_{max}.

Practical Application

In practice, the approach proposed here will be carried out in GWAS with possibly large numbers of interdependent SNPs. Also, the fact that test SNPs are ascertained at stage 1 for having marginal association effects may impact significance levels at stage 2 [15]. Thus, we advocate that both stages (finding the test SNP, testing the target SNP) be evaluated in a permutation framework, which will yield proper significance levels, that is, the type 1 error will be controlled for stages 1 and 2 combined. An additional benefit of permutation tests is that they automatically take care of multiple testing in GWAS and of the linkage disequilibrium between SNPs. Software (*sumstat* program) for these calculations has been made generally available (test statistic code 20, http://linkage.rockefeller.edu/ott/sumstat.html) and computes genome-wide p values corrected for multiple testing.

Application

To demonstrate our approach on a well-known dataset with established results, we applied the T_{sum} statistic to case-control data on age-related macular degeneration (AMD) [37, 38]. Initially, we analyzed the first of these two datasets but did not find any significant SNPs when conditioned on the genotypes of the three most significant SNPs (results not shown). This finding is in agree-

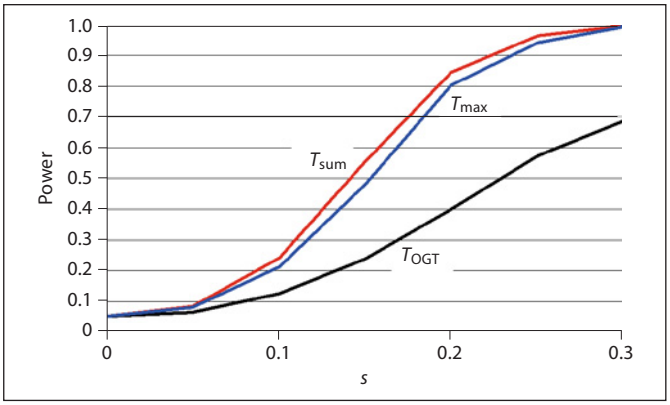


Fig. 2. Power (y-axis) as a function of a parameter *s* (x-axis) in a logistic regression model, where *s* is a measure of target SNP main and interaction effects.

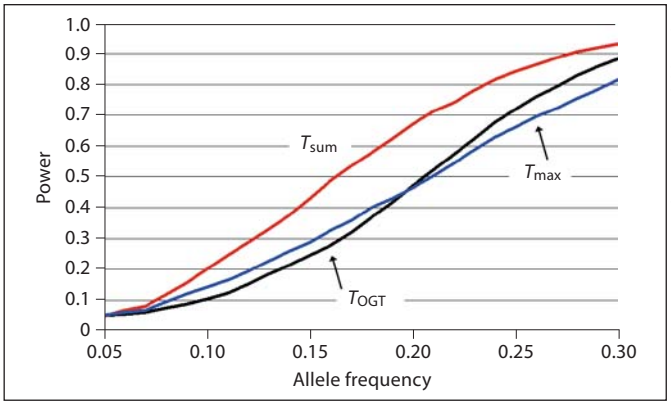


Fig. 3. Power (y-axis) as a function of the allele frequency (x-axis) of the *A* locus in Marchini et al.'s [13] multiplicative interaction model.

ment with a previous observation of no significant interactions among SNPs in this dataset [19]. Thus, we turned to the second of these two studies, that is, AMD data collected in Hong Kong [38]. After our quality control procedures, this dataset comprised 81,934 SNPs genotyped on 96 case and 127 control individuals. Previous analysis resulted in a significantly disease-associated SNP, rs10490924, whose functional significance had been established experimentally [38].

For our conditional analyses, χ^2 statistics were computed for the 2 × 3 genotype table of each SNP. After ordering the SNPs by decreasing χ^2, the best few SNPs were considered test SNPs. For a given test SNP, the T_{sum} statistic was carried for each of the 81,933 SNPs in the dataset and the largest of these among all SNPs was taken to be our genome-wide test statistic. This whole procedure was repeated in 10,000 randomization samples (*sumstat* program; labels *case* and *control* randomly permuted). The associated empirical significance level, p, was given by the proportion of randomization samples whose largest T_{sum} statistic was at least as large as the largest observed T_{sum}.

We selected a small set of test SNPs based on their χ^2 value. Table 4 shows the four best SNPs with $p < 0.30$ in the χ^2 test at stage 1 (the next ranked SNP has p = 0.58). For a conditional analysis as proposed here, there is no need for test SNPs to be significant; here we chose the most significant ones as this seems to be the most plausible approach to pick SNPs with disease involvement. For each of these four SNPs, one analysis was carried out in which the given SNP served as the test SNP and was tested against each of the variants (target SNPs) in the dataset.

Table 4. The four SNPs with p values less than 0.30 (genotype test) in a genome-wide screen for variants associated with age-related macular degeneration [38]

Rank	SNP	Chrom.	Position	χ^2	p
1	rs10490924	10	124204438	50.9622	0.0001
2	rs10504152	8	54292668	31.8137	0.0065
3	SNP_A-1702501	0	0	25.2921	0.1880
4	rs10520462	4	182400252	24.6355	0.2503

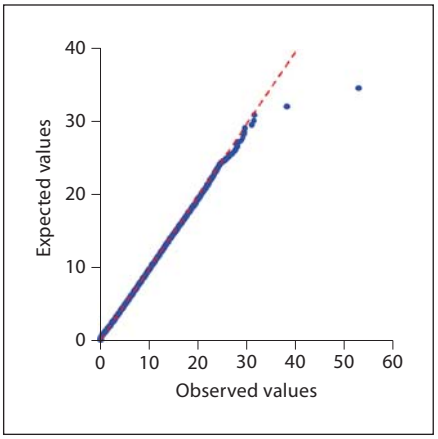
For several of the four test SNPs, highly significant results occurred with rs10490924 as the target SNP. However, this finding simply reflects the known strong main effect of rs10490924 and was disregarded. Two of the test SNPs furnished significant results (table 5): SNP_A-1702501 as the test SNP resulted in $T_{sum} = 37.93$ with rs1957491 as the target SNP (p = 0.0557), and rs10520462 as the test SNP resulted in $T_{sum} = 37.42$ with rs1599796 as the target SNP (p = 0.0179), where the reported p values are specific for one test SNP and incorporate the whole procedure of finding that test SNP at stage 1 and testing a target SNP at stage 2. The evidence for association with a disease of the two new SNPs, rs1957491 and rs1599796, is not strong but still significant or at least close to it. As the probability plot (carried out with *Systat* software) in figure 4 shows, despite the modest p value of 0.0557, at least two target SNPs exhibit unusually large values for SNP_A-1702501 as the test SNP. At any rate, the main effects in terms of χ^2 from the 2 × 3 genotype tables of these two SNPs are low (p = 0.18 and 1.00, respectively). Thus, these SNPs could not have been

Table 5. Two SNPs from table 4 serving as test SNPs and their best target SNPs

Test SNP	Target SNP					
	name	chrom.	position	T_{sum}	p	p main
SNP_A-1702501	rs1957491	14	43379465	37.9323	0.0557	0.1809
rs10520462	rs1599796	3	120726624	37.4237	0.0179	1.0000

found by single-locus analysis, but they appear to be associated with a disease through interaction with one of the four strongest SNPs in this analysis. Table 2 shows observed numbers of genotypes at SNPs SNP_A-1702501 (test SNP) and rs1957481 (target SNP). In a genome-wide search, the target SNP main effect of $\chi^2 = 12.8013$ (2 d.f.) is too small to be detected.

It is not the purpose of this paper to discuss AMD genetics and the functional relevance of the two suggestive SNPs found here. However, at least one of them, rs1599796, may well have a real effect. It is located on chromosome 3q13.33 in C3orf1, which has been described as encoding a membrane protein and showing generalized expression in all tissues [39].



Color version available online

Fig. 4. Probability plot (Q-Q plot) for all values of T_{sum} with SNP_A-1702501 as the test SNP in the AMD dataset [38]; observed values are on the x-axis while expected values, based on the χ^2 distribution with 6 d.f., are plotted on the y-axis.

Discussion

We made use of the concept of finding new loci anywhere in the genome, well-known in linkage analysis, given previously established disease-associated loci, and extended this approach to human case-control association studies. Of course, instead of conditioning on established loci, researchers may want to condition on candidate variants whether or not they show strong association with a disease. Our approach is analogous to the two-stage strategy proposed by Marchini et al. [13] in that loci detected at stage 1 do not themselves need to be significant, but our approach is different in that, at stage 2, we use a simple conditional test statistic instead of a full logistic regression analysis.

In addition to references already mentioned, several other papers have recommended two-stage procedures. Specifically, Kooperberg and Leblanc [40] focus on gene-gene interactions between SNPs with some marginal effects. Also, gene-environment approaches have been developed [41] that in the first stage select specific environmental effects and, in the second stage, carry out conditional tests similar to the ones proposed here.

Our favored test statistic is the sum, T_{sum}, over χ^2 values with 2 d.f. for the three genotypes at a given test locus. Of course, various refinements may be possible here. For example, association tests more powerful than the standard χ^2 genotype test have been described [42, 43] and could be implemented in connection with our approach. Also, an allelic form of these methods would condition on the three genotypes of the test SNP but then compute χ^2 for 2×2 tables of alleles, so that the sum of χ^2 over the three test SNP genotypes would have 3 d.f. However, the main purpose of our paper has been to demonstrate the usefulness of the principle of conditional search rather than fleshing out specific refinements. Because epistatic effects are likely to be ubiquitous [44], conditional approaches like the one suggested here are expected to be extremely useful in GWAS.

It should be pointed out that 'Interaction' is a concept that has been defined and interpreted in different ways. Two important classes are *essential* and *removable* inter-

actions, where the latter are robust to changes in scale [45]. Recently, methods have been developed to search for essential interactions on a genome-wide scale [46].

As mentioned above, χ^2 analyses leading to T_{sum} may also be carried out with logistic regression analysis. However, in samples of small sizes, like the ones discussed here, logistic regression analysis may fail (the iterative maximum likelihood estimation may not converge). While χ^2 analyses are also then unreliable when p values are obtained under large-sample assumptions (based on tables of the χ^2 distribution), permutation tests do not rely on distributional assumptions of the test statistic. In larger samples, the main advantage of logistic regression analysis is that risk factors other than SNP genotypes may be allowed for.

As pointed out by an anonymous reviewer, our approach (and power calculation) assumes that the test SNP is correctly identified in the stage 1 analysis. Of course, it is possible that the test SNP reflects a false-positive result, particularly when its significance is only marginal. However, when both stages 1 and 2 are evaluated in randomization samples (permutation testing) then any uncertainty as to the validity of the test SNP will be reflected in the overall significance level (see Practical Application).

Acknowledgments

Grant support from the China Natural Science Foundation (to Y.Y. and J.O.) is gratefully acknowledged. This work was also supported in part by NIH grants AG026916 and HL084410.

Electronic Resources

Sumstat program: http://www.big.ac.cn/genemapping/ or http://linkage.rockefeller.edu/ott/.

Systat software: http://www.systat.com.

AMD dataset: http://variation.yale.edu/dataDownload.html.

References

1 McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–369.
2 Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P, Sullivan PF: Genomewide association studies: history, rationale, and prospects for psychiatric disorders. Am J Psychiatry 2009;166:540–556.
3 Dupuis J, Brown PO, Siegmund D: Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 1995;140:843–856.
4 Lander ES, Botstein D: Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 1986;83:7353–7357.
5 Ott J, Falk CT: Epistatic association and linkage analysis in human families. Hum Genet 1982;62:296–300.
6 Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farrall M, Barnett AR, Baln SC, Todd JA: A genome-wide search for human type 1 diabetes susceptibility genes. Nature 1994;371:130–136.
7 Hashimoto L, Habita C, Beressi JP, Delepine M, Besse C, Cambon-Thomsen A, Deschamps I, Rotter JI, Djoulah S, James MR, Froguel P, Weissenbach J, Lathrop GM, Julier C: Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. Nature 1994;371:161–164.
8 Shah SH, Schmidt MA, Mei H, Scott WK, Hauser ER, Schmidt S: Searching for epistatic interactions in nuclear families using conditional linkage analysis. BMC Genet 2005; 6(suppl 1):S148.
9 Qiao Q, Osterholm AM, He B, Pitkaniemi J, Cordell HJ, Sarti C, Kinnunen L, Tuomilehto-Wolf E, Tryggvason K, Tuomilehto J: A genome-wide scan for type 1 diabetes susceptibility genes in nuclear families with multiple affected siblings in Finland. BMC Genet 2007;8:84.
10 Angquist L, Hossjer O, Groop L: Strategies for conditional two-locus nonparametric linkage analysis. Hum Hered 2008;66:138–156.
11 Di Y, Thompson EA: Conditional tests for localizing trait genes. Hum Hered 2009;68: 139–150.
12 Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J: Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Ann Hum Genet 2000;64:413–417.
13 Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37:413–417.
14 Hoh J, Wille A, Ott J: Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 2001;11: 2115–2119.
15 Long Q, Zhang Q, Ott J: Detecting disease-associated genotype patterns. BMC Bioinformatics 2009;10:S75.
16 Manolio TA, Collins FS: Genes, environment, health, and disease: facing up to complexity. Hum Hered 2007;63:63–66.
17 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: Finding the missing heritability of complex diseases. Nature 2009;461:747–753.
18 Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 2003;56:73–82.
19 Zhang Y, Liu JS: Bayesian inference of epistatic interactions in case-control studies. Nat Genet 2007;39:1167–1173.
20 Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ: Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 2008;4:e1000130.
21 Wu TT, Chen YF, Hastie T, Sobel E, Lange K: Genomewide association analysis by lasso penalized logistic regression. Bioinformatics 2009;28:28.

22 Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH: Reducing dimensionality for prediction of genome-wide breeding values. Genet Sel Evol 2009;41:29.
23 Long N, Gianola D, Rosa GJ, Weigel KA, Avendano S: Comparison of classification methods for detecting associations between SNPs and chick mortality. Genet Sel Evol 2009;41:18.
24 Bellone G, Carbone A, Busso V, Scirelli T, Buffolino A, Smirne C, Novarino A, Bertetto O, Tosetti L, Emanuelli G: Antagonistic interactions between gemcitabine and 5-fluorouracil in the human pancreatic carcinoma cell line capan-2. Cancer Biol Ther 2006;5:1294–1303.
25 Descot A, Hoffmann R, Shaposhnikov D, Reschke M, Ullrich A, Posern G: Negative regulation of the EGFR-MAPK cascade by actin-MAL-mediated Mig6/Errfi-1 induction. Mol Cell 2009;35:291–304.
26 Assmann G, Voswinkel J, Mueller M, Bittenbring J, Koenig J, Menzel A, Pfreundschuh M, Roemer K, Melchers I: Association of rheumatoid arthritis with Mdm2 SNP309 and genetic evidence for an allele-specific interaction between MDM2 and p53 P72R variants: a case control study. Clin Exp Rheumatol 2009;27:615–619.
27 Carrington M, Dean M, Martin MP, O'Brien SJ: Genetics of HIV-1 infection: chemokine receptor CCR5 polymorphism and its consequences. Hum Mol Genet 1999;8:1939–1945.
28 Kaur G, Mehra N: Genetic determinants of HIV-1 infection and progression to AIDS: susceptibility to HIV infection. Tissue Antigens 2009;73:289–301.
29 Cordell HJ: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 2002;11:2463–2468.
30 Frankel WN, Schork NJ: Who's afraid of epistasis? Nat Genet 1996;14:371–373.
31 Culverhouse R, Suarez BK, Lin J, Reich T: A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 2002;70:461–471.
32 Thomson G, Valdes AM: Conditional genotype analysis: detecting secondary disease loci in linkage disequilibrium with a primary disease locus. BMC Proc 2007;1(suppl 1):S163.
33 Thomson G, Barcellos LF, Valdes AM: Searching for additional disease loci in a genomic region. Adv Genet 2008;60:253–292.
34 Armitage P, Berry G, Matthews JNS: Statistical Methods in Medical Research, ed 4. Malden, Blackwell Science, 2002.
35 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.
36 Snedecor GW, Cochran WG: Statistical Methods, ed 8. Ames, Iowa State University Press, 1989.
37 Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–389.
38 Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, Tam PO, Chan WM, Lam DS, Snyder M, Barnstable C, Pang CP, Hoh J: HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 2006;314:989–992.
39 Escarceller M, Pluvinet R, Sumoy L, Estivill X: Identification and expression analysis of C3orf1, a novel human gene homologous to the Drosophila RP140-upstream gene. DNA Seq 2000;11:335–338.
40 Kooperberg C, Leblanc M: Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genet Epidemiol 2008;32:255–263.
41 Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ: Exploiting gene-environment interaction to detect genetic associations. Hum Hered 2007;63:111–119.
42 Zheng G, Freidlin B, Gastwirth JL: Comparison of robust tests for genetic association using case-control studies. IMS Lecture Notes-Monograph Series 2006;49:253–265.
43 Matthews AG, Haynes C, Liu C, Ott J: Collapsing SNP genotypes in case-control genome-wide association studies increases the type 1 error rate and power. Stat Appl Genet Mol Biol 2008;7:Art. 23.
44 Moore JH, Williams SM: Epistasis and its implications for personal genetics. Am J Hum Genet 2009;85:309–320.
45 Breslow NE, Day NE: The analysis of case-control studies. Lyon, International Agency of Cancer Research, 1980.
46 Wu C, Zhang H, Liu X, DeWan A, Dubrow R, Ying Z, Yang Y, Hoh J: Detecting essential and removable interactions in genome-wide association studies. Statistics and Its Interface 2009;2:161–170.