European Journal of Human Genetics (2010) 18, 960–964
© 2010 Macmillan Publishers Limited All rights reserved 1018-4813/10
www.nature.com/ejhg

ARTICLE

Detection of susceptibility genes as modifiers due to subgroup differences in complex disease

Sarah E Bergen[1,2], Brion S Maher[1,2,3], Ayman H Fanous[2,3,4,5] and Kenneth S Kendler[1,2,3]

Complex diseases invariably involve multiple genes and often exhibit variable symptom profiles. The extent to which disease symptoms, course, and severity differ between affected individuals may result from underlying genetic heterogeneity. Genes with modifier effects may or may not also influence disease susceptibility. In this study, we have simulated data in which a subset of cases differ by some effect size (ES) on a quantitative trait and are also enriched for a risk allele. Power to detect this 'pseudo-modifier' gene in case-only and case–control designs was explored blind to case substructure. Simulations involved 1000 iterations and calculations for 80% power at $P<0.01$ while varying the risk allele frequency (RAF), sample size (SS), ES, odds ratio (OR), and proportions of the case subgroups. With realistic values for the RAF (0.20), SS (3000) and ES (1), an OR of 1.7 is necessary to detect a pseudo-modifier gene. Unequal numbers of subjects in the case groups result in little decrement in power until the group enriched for the risk allele is $<30\%$ or $>70\%$ of the total case population. In practice, greater numbers of subjects and selection of a quantitative trait with a large range will provide researchers with greater power to detect a pseudo-modifier gene. However, even under ideal conditions, studies involving alleles with low frequencies or low ORs are usually underpowered for detection of a modifier or susceptibility gene. This may explain some of the inconsistent association results for many candidate gene studies of complex diseases.

European Journal of Human Genetics (2010) **18,** 960–964; doi:10.1038/ejhg.2010.39; published online 31 March 2010

Keywords: modifier; simulation; subtyping; complex disease

INTRODUCTION

As the risk genes for complex diseases are rapidly being identified,[1–3] there has been increasing attention to the factors influencing variability within these disorders. For complex genetic disorders in which multiple genes act in concert to produce the disease, variability in phenotypic expression seems likely to result, at least in part, from underlying genetic heterogeneity. Genes impacting age of onset, severity, and differences in symptom clusters, but not susceptibility to illness have been termed modifier genes.[4] Several such genes have now been identified for diseases such as sickle cell anemia,[5] cystic fibrosis,[6] nonsyndromic cleft lip,[7] and spinocerebellar ataxia type II.[8,9]

Although susceptibility genes – those which influence disease liability – and modifier genes – those which impact on clinical variation within the illness – can be distinct, it is also possible for one gene to predispose individuals to a disease as well as specific symptom dimensions within the illness.[4] Such 'susceptibility-modifier' genes have been identified for schizophrenia (dysbindin,[10–12] COMT,[13–15] and DISC1[16,17]), cystic fibrosis (CFTR),[18] ataxia-telangiectasia (ATM),[19] long-QT syndrome (NOS1AP),[20] and type 2 diabetes (IRS1).[21]

Difficulty in replicating findings for association studies of complex diseases has led to the stratification of subjects based on variation in disease characteristics as one plausible way to enhance the signal to noise ratio by reducing heterogeneity. Consequently, there have been attempts to identify biological features specific to particular forms of illness. However, classifying clinical subtypes of complex diseases has proven exceedingly difficult. Age of onset, severity, and familial or sporadic inheritance patterns have sometimes been used to subdivide case populations. Other strategies for grouping patients have focused on symptomatology, either through use of *a priori* criteria or by the more statistically rigorous methods of cluster or latent class analysis. Subjects can then be categorized as high or low scorers for a given factor. There are also methods that allow for the maximization of evidence for association by covariate-based subdividing without *a priori* cut points or data processing.[22,23] Each of these categorization approaches has some appeal, but the best approach for each complex disease has yet to be determined.

As genes that confer susceptibility to a form of illness with a distinct symptom profile would manifest as susceptibility-modifier genes, the results of modifier gene association studies may yield information regarding variation in the genetic architecture of complex disease liability in addition to variability in symptom expression.

We suggest there are two particularly plausible mechanisms whereby a gene variant is associated with a symptom in a complex disease. First, the disorder is etiologically homogeneous and this gene 'truly' impacts on that symptom – a true modifier. Second, the disorder is etiologically heterogeneous. This 'pseudo-modifier' gene is really a risk gene but only for one subtype AND the subtypes differ on the levels of this particular symptom. We term this type of gene a 'pseudo-modifier' because its effects on the symptoms in question actually arise from it conferring liability to a particular disease subtype. In this paper, we study this second mechanism to see under what circumstances it might be detected.

[1]Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA; [2]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA; [3]Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA; [4]Washington VA Medical Center, Washington, DC, USA; [5]Department of Psychiatry, Georgetown University Medical Center, Washington, DC, USA
Correspondence: Dr SE Bergen, Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetics Research, Massachusetts General Hospital, 185 Cambridge St, 6th floor, Boston, MA 02114, USA. Tel: +1 617 643 0992; Fax: +1 617 726 0830; E-mail: sbergen@gmail.com

Received 10 December 2008; revised 8 December 2009; accepted 5 February 2010; published online 31 March 2010

To do so, we simulated two case groups, for one of which the gene variant influencing symptom variability also confers disease susceptibility. The other case group arrives at the disease state through another, unspecified mechanism. A control group was simulated as well, but as power for case–control (susceptibility) analyses has been thoroughly investigated elsewhere, these results are included here for comparison purposes only. Case-only designs (for modifier effects) were considered, blind to case substructure, as risk allele frequency (RAF), sample size (SS), odds ratio (OR), effect size (ES), and proportion of cases with the pseudo-modifier allele were varied.

METHODS

Two case groups and a control group were simulated according to a range of specified parameters, then tested for power to detect the pseudo-modifier gene of interest. Simulations were carried out using the software program SAS 9.1 or 9.1.3.[24] All sets of simulations involved 1000 iterations and calculations of power given the RAF, OR, SS, and ES. We created two case subgroups differentiated on mean group differences for an unspecified, normally distributed, quantitative trait. The type II cases were enriched for the pseudo-modifier allele of interest, whereas the type I cases were not. Figure 1 illustrates the two case population distributions and their combined distribution when subgroup membership is unknown.

We did not directly simulate an effect of the variant on the quantitative trait. Instead, we simulated a variant with population allele frequency in controls and type I cases and RAF*OR in type II cases. Importantly, this results only in an increased RAF among the type II cases. In case–control comparisons where the number of cases and controls is equal, the effective OR is then 1+((OR−1)/2). For example, a risk allele with a frequency of 0.1 at an OR of 1.4 would yield frequencies of 0.1, 0.1, and 0.14 in controls, type I cases, and type II cases, respectively, in a sufficiently large sample. In case–control comparisons, grouping the heterogeneous case sets, the allele frequencies would be 0.1 in controls and 0.12 in cases.

For use in case-only analyses, a quantitative phenotype was simulated sampling from a normal distribution with a mean of 0 and standard deviation of 1 in type I cases and a mean equal to the ES, the standardized mean difference of a trait between the two groups of cases, in type II cases, thus indirectly creating an association between the SNP and the quantitative trait. That is, the case group enriched for the risk allele also has a mean difference from the other case group.

For each set of parameter specifications, power to detect the influence of the impact of genotype on variation in the combined case groups (case-only) was calculated as well as power to detect the allele as conferring disease susceptibility (case–control). Although the case–control power calculations are not novel, they provide a useful comparison for the case-only investigations. We chose to use a one-stage design as studies exploring modifier gene influences are not contingent on susceptibility gene association results.

Main analyses

All analyses were performed blind to case type. Unless otherwise specified, simulations included 3000 total subjects in which 750 were type I cases, 750 were type II cases, and 1500 were controls. RAFs were varied from 0.10 to 0.50, and ORs of 1.1–2.0 by increments of 0.10 were modeled, initially holding the ES at 1. Regression tests using PROC REG in SAS were conducted for the case-only modifier analyses, and a Cochran Armitage trend test was used for the case–control susceptibility analyses.[24] We defined sufficient power for detection as 0.8 or greater with a *P*-value ≤ 0.01.

Additional analyses

The RAF was fixed at 0.20 and ES at 1 for analyses in which other parameters varied. To explore the effects of SS, the total number of subjects was increased incrementally by 1000 from 2000 to 6000 while maintaining the same proportion of subjects in each group. Additionally, as it is implausible that two subpopulations of affected individuals would naturally divide the subject pool neatly in half, we also varied the percent of Type II subjects, possessing an enriched proportion of the risk-conferring allele. Total case and control numbers were held even. Furthermore, the ESs that might be observed could vary considerably and depend entirely on the phenotypic trait assessed. We consequently modeled a broad range of ESs from 0.5 to 3.0 with increments of 0.5 representing mean differences of half a standard deviation to three full standard deviations.

RESULTS

To detect modifier associations that result from underlying genetic heterogeneity, in which the allele impacts on disease risk in only one subgroup, with our core set of parameter specifications (SS of 3000, ES of 1, and RAF of 0.20) we had sufficient power to detect a pseudo-modifier association with an OR of 1.7 or greater (Figure 2). For comparison, a susceptibility allele with an OR of 1.2 is detectable under the same conditions.

Allele frequency

RAFs of 0.10 or less are not sufficient to detect pseudo-modifier genes with a one standard deviation mean trait difference.

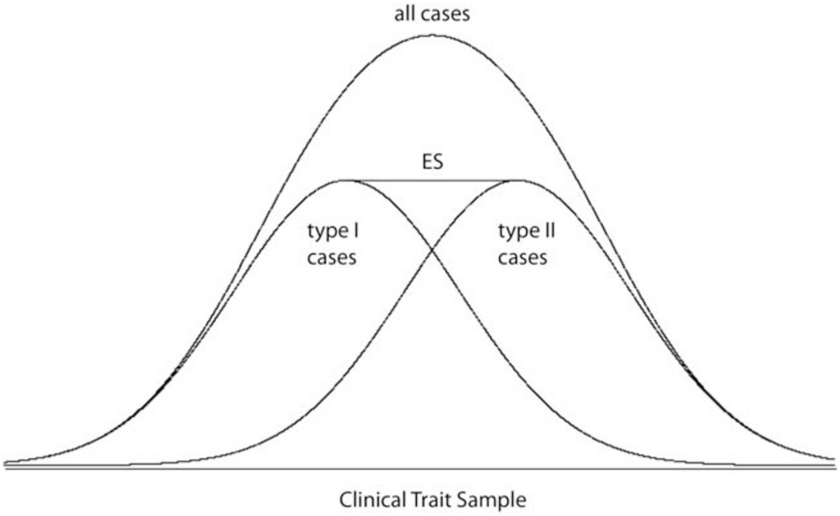

Figure 1 Case population distributions in relation to a clinical trait scale. Type I cases are depicted as scoring lower on the scale. Type II cases, enriched for the pseudo-modifier allele of interest, score higher by an effect size (ES) difference of one standard deviation in most simulations. The combined case population is also shown because investigators (and our analyses) are blind to case substructure.

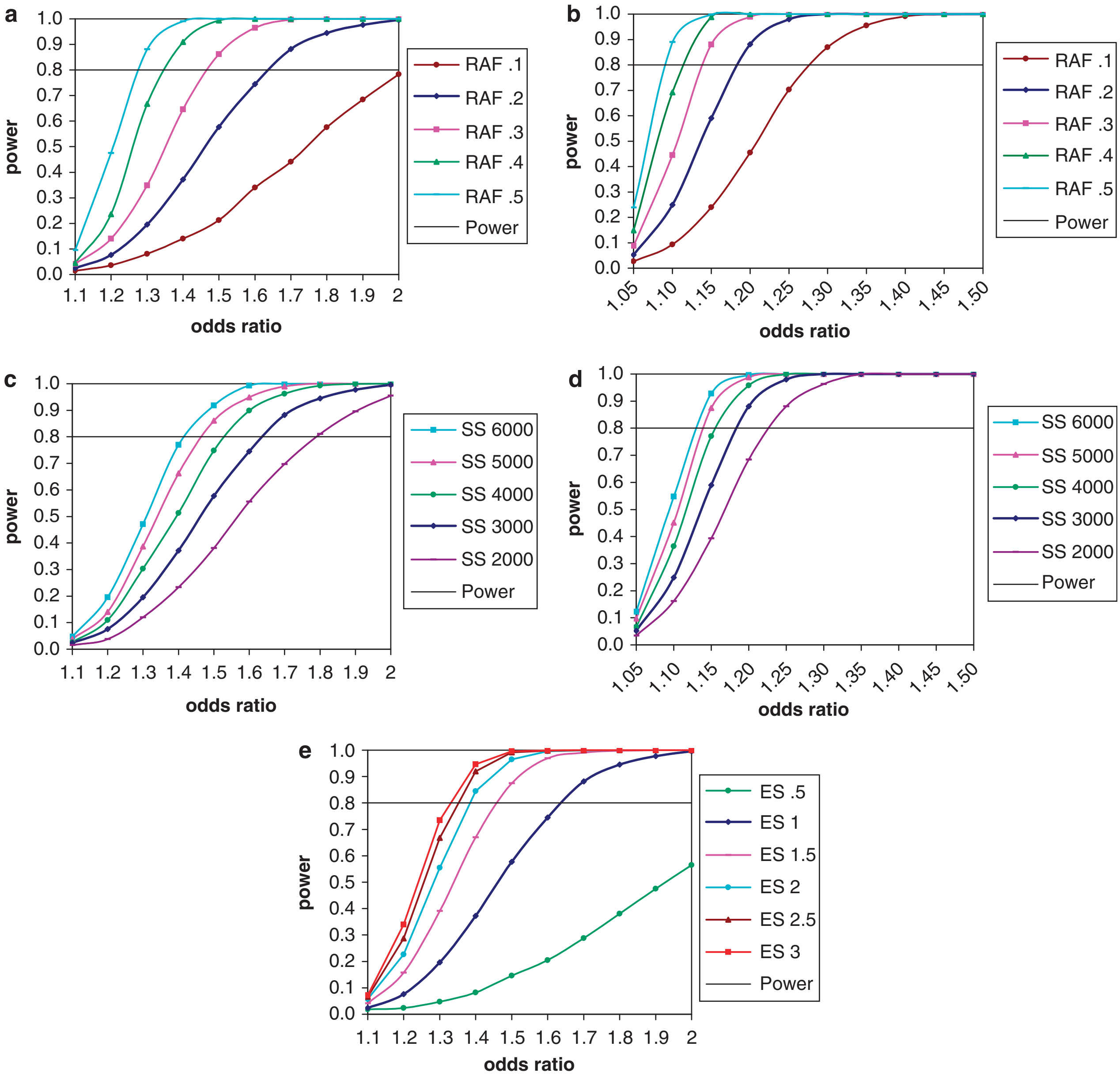

Figure 2 Power by odds ratio (OR) curves. When fixed, the risk allele frequency (RAF)=0.2, effect size (ES)=1, and sample size (SS)=3000. Lines with these parameter specifications are represented on each plot and emphasized for frame of reference. (**a**) Power to detect pseudo-modifier genes for RAF of 0.1–0.5. (**b**) Power to detect a susceptibility gene in case–control analyses varying the RAF from 0.1 to 0.5 and OR from 1.05 to 1.50. (**c**) Power for pseudo-modifier gene detection for SS=2000–6000. (**d**) Power for susceptibility gene detection with SS=2000–6000. (**e**) ES 0.5–3.0 impact on power to detect pseudo-modifier effects.

However, susceptibility gene detection is possible with an OR of 1.3 or greater. Power curves grow progressively steeper as the RAF increases, culminating with detection of OR 1.15 and 1.3 for the susceptibility and pseudo-modifier gene analyses, respectively.

Sample size

With 2000 total subjects, only ORs of 1.8 and greater are sufficient for detection of pseudo-modifier effects. However, each additional 1000 subjects lowers the detectable OR by ~0.10 until an SS of 6000 is used. With this large SS, power to detect an OR of 1.4 is just under 0.80 but falls sharply to ~0.50 for OR detection of 1.3. Case–control simulations for susceptibility gene effects exhibit much steeper curves with a smaller range, allowing for detection of an RAF of 0.20 or greater and OR of 1.15–1.25 across all SSs examined.

Effect size

When the mean phenotypic differences between subgroups is less than half a standard deviation, detecting pseudo-modifier effects is unlikely with ORs under two. Standard deviation differences of 1 and 1.5 are distinguishable at ORs of 1.7 and 1.5, respectively. From standard deviations of 2–3, however, the increased phenotypic disparity does not confer markedly enhanced detectability for pseudo-modifier genes.

Unbalanced case groups

Little decrement in power for the detection of pseudo-modifier effects was observed across a broad range of the percent of subjects with the subtype containing an enriched proportion of the risk-conferring allele. At any OR for which there is sufficient power for detection,

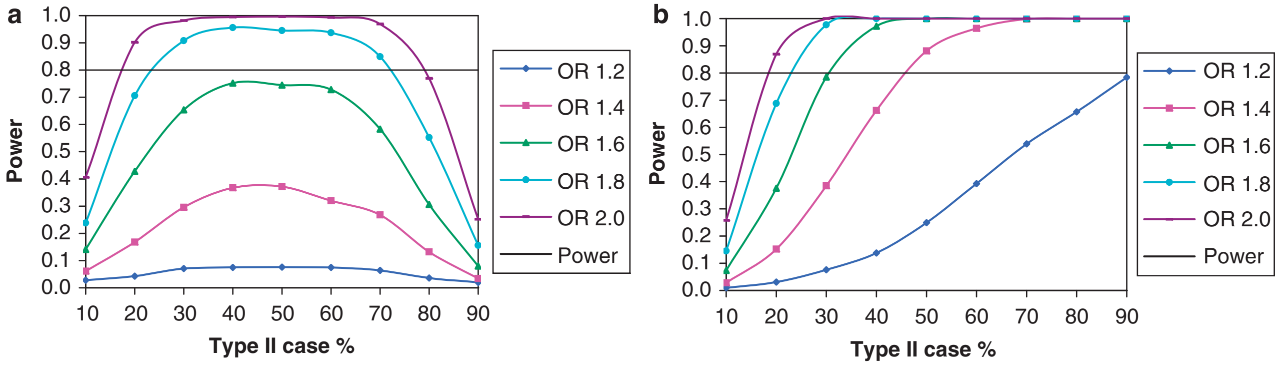

Figure 3 Effects of varying the proportion of Type II cases, enriched for the risk allele, on power across odds ratios (OR) 1.2–2.0. (**a**) Case-only results for pseudo-modifier discrimination. (**b**) Power for risk allele detection in case–control analyses.

30–70% of the Type II cases are sufficient, and for high ORs, an even broader range may suffice (Figure 3).

DISCUSSION

From these results, it is clear that the discovery of modifier gene effects that arise from genetic heterogeneity in complex diseases is critically dependent on most of the parameter estimates examined here. However, it is interesting to note that the proportion of cases with the risk-conferring allele can vary between 30 and 70% with little observed deviation in power. As subtypes are unlikely to evenly divide an affected population, this is a reassuring finding. On the other hand, RAF is a vital factor in detecting pseudo-modifier genes. With 3000 subjects, an RAF of 0.10 (or less) is insufficient for detection of a pseudo-modifier gene for any OR under 2.0, and even under ideal circumstances with an RAF of 0.50, the OR must be 1.3 or greater.

ES is another important determinant of pseudo-modifier gene detection, although the phenotypic differences between case groups have no impact on susceptibility gene detection. This is an important consideration when selecting a trait on which to explore modifier effects as the groups should minimally differ by one standard deviation and preferably two or more on the trait of interest. As group membership is generally unknown, high trait variance is the best selection criterion usually available.

In practice, SS is the most manipulable of the parameters explored here. It is intuitively obvious that greater numbers of subjects confer greater power, but beyond 5000, the additional effort of subject recruitment and assessment may not yield sufficiently enhanced power to be worthwhile. Even with very high numbers of subjects, detection of genes with low RAFs or low ORs is extremely difficult. For many complex diseases this may explain conflicting results from association studies.

For the parameters tested, attempts were made to examine realistic values. For example, the Affymetrix Mapping 500K Array and GenomeWide Human Mapping 5.0 Array both report average minor allele frequencies of 0.22.[25] and the Illumina HumanHap550 and 650Y detect SNPs with average minor allele frequencies of 0.20–0.23 depending on the population sampled.[26] Consequently, simulations in which risk allele frequencies were held constant were given values of 0.20. Moreover, the OR range we used is comparable to ORs commonly reported for complex genetic diseases[1,2,27,28] (eg ∼1.1–2.0).

The added power conferred by greater subject numbers has led to recent increases in multi-center collaborations generating cohorts consisting of thousands of individuals. Accordingly, the number of subjects used for most simulations in this study approximated that of some of these cooperative efforts. These include several of the Genetic Association Information Network (GAIN) studies such as that for ADHD (involving 2877 participants), major depression (3720), bipolar disorder (3316), schizophrenia (5189), and psoriasis (2902).[29] The Wellcome Trust Consortium is another large collaborative effort, which examined 2000 cases for each of 7 major complex diseases and a shared set of 3000 controls.[1]

Despite our attempts at modeling realistic values, some limitations were imposed out of necessity. The simulations presented here only assess the impact of a single gene on the predisposition to a certain form of a complex disease. In fact, a more likely scenario involves overlapping constellations of susceptibility genes as well as environmental insults that are also not included in these models. In addition, we have presumed that the minor allele is the risk-conferring allele. However, common alleles with small main effects may act in concert with alleles in other genes to additively or epistatically influence liability to complex diseases, and we have not modeled these possibilities. Furthermore, the simulations presented here were restricted to two subtypes, when in reality, many more subtypes may exist for some diseases.

We have additionally presumed that the modifier effects of the allele in question are restricted to (or only assessed in) the case population. Detection of more subtle expression in unaffected individuals, when possible, might allow for greater power to detect the allele.[4] Gene detection then hinges not only on disease expression but degrees of symptom expression as well, drawing on increased information to yield enhanced power.

The results presented here are applicable to nearly every complex genetic disease for which subtypes may exist. Several diseases may manifest with convergent symptom profiles but arise through different etiological mechanisms. The extent to which subpopulations differ in their symptoms can yield clues to underlying biological differences. For example, diabetes has two main subtypes (I and II) both of which present with high blood glucose levels and similar symptoms such as extreme thirst, fatigue, and blurred vision. However, in type I diabetes, the symptoms are due to the destruction of insulin-producing cells, whereas type II diabetes occurs when the tissues become resistant to insulin or produce too little. Later age of onset and higher body weight are linked to, but not diagnostic of, type II diabetes.[30] These subtle phenotypic differences indicative of very distinct pathophysiological origins are precisely the type of clues sought to unlock the etiology of complex diseases.

Association studies examining modifier effects may actually uncover subtype-specific susceptibility genes. Whether variable symptom profiles for numerous diseases are due to modifier genes and environmental influences, differing underlying genetic architecture, or some combination of these possibilities will likely remain under investigation for many years to come. The simulation results presented here indicate there is reasonable power to detect pseudo-modifier genes under favorable conditions such as a high ES, OR, and RAF, but they may well be missed under less ideal circumstances. These results can be used to inform researchers as to the relative power for studies of complex disease under a range of conditions when symptom variation is due to different genetic subtypes.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was funded by NIH grants MH41953 to KSK, DE016632 to BSM, training grant MH20030 supporting SEB, and a VA Merit Review Program grant to AHF.

1 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.

2 Barrett JC, Hansoul S, Nicolae DL *et al*: Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 2008; **40**: 955–962.

3 Jakobsdottir J, Conley YP, Weeks DE, Mah TS, Ferrell RE, Gorin MB: Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am J Hum Genet* 2005; **77**: 389–407.

4 Fanous AH, Kendler KS: Genetic heterogeneity, modifier genes, and quantitative phenotypes in psychiatric illness: searching for a framework. *Mol Psychiatry* 2005; **10**: 6–13.

5 Steinberg MH: Predicting clinical severity in sickle cell anaemia. *Br J Haematol* 2005; **129**: 465–481.

6 Salvatore F, Scudiero O, Castaldo G: Genotype-phenotype correlation in cystic fibrosis: the role of modifier genes. *Am J Med Genet* 2002; **111**: 88–95.

7 Zucchero TM, Cooper ME, Maher BS *et al*: Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *N Engl J Med* 2004; **351**: 769–780.

8 Hayes S, Turecki G, Brisebois K *et al*: CAG repeat length in RAI1 is associated with age at onset variability in spinocerebellar ataxia type 2 (SCA2). *Hum Mol Genet* 2000; **9**: 1753–1758.

9 Pulst SM, Santos N, Wang D *et al*: Spinocerebellar ataxia type 2: polyQ repeat variation in the CACNA1A calcium channel modifies age of onset. *Brain* 2005; **128**: 2297–2303.

10 Straub RE, Jiang Y, MacLean CJ *et al*: Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am J Hum Genet* 2002; **71**: 337–348.

11 Fanous AH, van den Oord EJ, Riley BP *et al*: Relationship between a high-risk haplotype in the DTNBP1 (dysbindin) gene and clinical features of schizophrenia. *Am J Psychiatry* 2005; **162**: 1824–1832.

12 DeRosse P, Funke B, Burdick KE *et al*: Dysbindin genotype and negative symptoms in schizophrenia. *Am J Psychiatry* 2006; **163**: 532–534.

13 de Chaldee M, Laurent C, Thibaut F *et al*: Linkage disequilibrium on the COMT gene in French schizophrenics and controls. *Am J Med Genet* 1999; **88**: 452–457.

14 McClay JL, Fanous A, van den Oord EJ *et al*: Catechol-O-methyltransferase and the clinical features of psychosis. *Am J Med Genet B Neuropsychiatr Genet* 2006; **141B**: 935–938.

15 DeRosse P, Funke B, Burdick KE *et al*: COMT genotype and manic symptoms in schizophrenia. *Schizophr Res* 2006; **87**: 28–31.

16 Thomson PA, Wray NR, Millar JK *et al*: Association between the TRAX/DISC locus and both bipolar disorder and schizophrenia in the Scottish population. *Mol Psychiatry* 2005; **10**: 657–668, 616.

17 DeRosse P, Hodgkinson CA, Lencz T *et al*: Disrupted in schizophrenia 1 genotype and positive symptoms in schizophrenia. *Biol Psychiatry* 2007; **61**: 1208–1210.

18 Dean M, Santis G: Heterogeneity in the severity of cystic fibrosis and the role of CFTR gene mutations. *Hum Genet* 1994; **93**: 364–368.

19 Verhagen MMM, Abdo WF, Willemsen MAAP *et al*: Clinical spectrum of ataxia-telangiectasia in adulthood. *Neurology* 2009; **73**: 430–437.

20 Crotti L, Cristina M, Insolia R *et al*: NOS1AP is a genetic modifier of the long-QT syndrome. *Circulation* 2009; **120**: 1657–1663.

21 Rung J, Cauchi S, Albrechtsen A, *et al*: Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature* 2009; **41**: 1110–1117.

22 Perdry H, Maher BS, Babron MC, McHenry T, Clerget-Darpoux F, Marazita ML: An ordered subset approach to including covariates in the transmission disequilibrium test. *BMC Proc* 2007; **1** (Suppl 1): S77.

23 Macgregor S, Craddock N, Holmans PA: Use of phenotypic covariates in association analysis by sequential addition of cases. *Eur J Hum Genet* 2006; **14**: 529–534.

24 SAS Institute Inc.: *SAS, Version 9.1.3*. Cary, NC: SAS Institute Inc., 2005.

25 Affymetrix Inc. 2006. 7–15–2008 http://www.affymetrix.com.

26 Illumina Inc. 2006. 7–16–2008 http://www.illumina.com.

27 Allen NC, Bagade S, McQueen MB *et al*: Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat Genet* 2008; **40**: 827–834.

28 Zeggini E, Scott LJ, Saxena R *et al*: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638–645.

29 Manolio TA, Rodriguez LL, Brooks L *et al*: New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet* 2007; **39**: 1045–1051.

30 National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Diabetes Information Clearinghouse (NDIC): Diabetes overview. *NIH* 2006. 7–12–2008 http://diabetes.niddk.nih.gov/dm/pubs/overview/index.htm.