Overrepresentation of rare variants in a specific ethnic group may confuse interpretation of association analyses

Dianne Keen-Kim¹, Carol A. Mathews³, Victor I. Reus³, Thomas L. Lowe³, Luis Diego Herrera⁴, Cathy L. Budman⁵, Varda Gross-Tsur⁶, Ann E. Pulver⁷, Ruth D. Bruun⁵, Gerald Erenberg⁸, Allan Naarden⁹, Chiara Sabatti² and Nelson B. Freimer¹,*

¹Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior and ²Departments of Human Genetics and Statistics, University of California, Los Angeles, CA, USA, ³Department of Psychiatry, University of California, San Francisco, CA, USA, ⁴Hospital Nacional de Niños, San Jose, Costa Rica, ⁵Movement Disorders Center in Psychiatry, North Shore University Hospital, Manhasset, NY, USA, ⁶Neuropediatric Unit, Shaare Zedek Medical Center, Jerusalem, Israel, ⁷Psychiatry and Behavioral Sciences at the Johns Hopkins University School of Medicine, Baltimore, MD, USA, ⁸Department of Neurology, Cleveland Clinic Foundation, OH, USA and ⁹Department of Neurology, University of Texas Southwestern Medical School, Dallas, TX, USA

Received July 14, 2006; Revised and Accepted October 5, 2006

Rare sequence variants may be important in understanding the biology of common diseases, but clearly establishing their association with disease is often difficult. Association studies of such variants are becoming increasingly common as large-scale sequence analysis of candidate genes has become feasible. A recent report suggested \textit{SLITRK1} (\textit{Slit} and \textit{Trk}-like 1) as a candidate gene for Tourette Syndrome (TS). The statistical evidence for this suggestion came from association analyses of a rare 3'-UTR variant, var321, which was observed in two patients but not observed in more than 2000 controls. We genotyped 307 Costa Rican and 515 Ashkenazi individuals (TS probands and their parents) and observed var321 in five independent Ashkenazi parents, two of whom did not transmit this variant to their affected child. Furthermore, we identified var321 in one subject from an Ashkenazi control sample. Our findings do not support the previously reported association and suggest that var321 is overrepresented among Ashkenazi Jews compared with other populations of European origin. The results further suggest that overrepresentation of rare variants in a specific ethnic group may complicate the interpretation of association analyses of such variants, highlighting the particular importance of precisely matching case and control populations for association analyses of rare variants.

INTRODUCTION

In gene mapping of rare diseases with Mendelian inheritance patterns, determining whether a particular variant is relevant to disease causation is generally straightforward; one observes that it segregates with the disease in pedigrees and is not observed in unaffected relatives or control subjects. Such is also the case for rare familial forms of common, complex disorders such as Alzheimer’s disease (1). In contrast, the complex inheritance patterns of most common diseases make it difficult to demonstrate convincingly that a particular variant segregates with the disease in pedigrees. Therefore, it is necessary to show the association of particular variants in cases compared with control subjects, in samples drawn from a population. Several factors complicate the interpretation of such studies. For example, analyses of rare variants lack power because it is difficult to ascertain large numbers of cases possessing the variant. Additionally, while rare

*To whom correspondence should be addressed at: UCLA Center for Neurobehavioral Genetics, Gonda Center, Rm. 3506, 695 Charles E. Young Dr S., Box 951761, Los Angeles, CA 90095, USA. Tel: +1 3107949571; Fax: +1 3107949613; Email: nfreimer@mednet.ucla.edu

© 2006 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
variants are usually identified by comprehensive screening of case samples, control samples are typically screened only for the variants identified in the case samples; this produces an ascertainment bias (2,3). We now present data for Tourette Syndrome (TS) that suggest that hidden population stratification due to the overrepresentation of rare variants within particular ethnic populations may be a factor that particularly complicates the interpretation of association analyses of rare variants.

TS is a movement disorder with childhood onset, characterized by vocal and motor tics and often accompanied by neuropsychiatric symptoms such as obsessive-compulsive behaviors, attentional problems or impulsivity. TS is highly heritable but is considered genetically complex (4,5). Abelson et al. (6) recently proposed SLITRK1 (Slit and Trk-like 1) as a TS candidate gene. SLITRK1 maps about 350 kb from the breakpoint of a balanced de novo chromosome 13 inversion found during prenatal diagnosis of a child who later developed TS. In 174 TS patients of European ancestry, they found one SLITRK1 coding frameshift mutation (one patient, varCDfs) and one non-coding variant in the SLITRK1 untranslated region (two patients, var321), but did not find these variants in 1800 (varCDfs) and 2148 (var321) controls, also of European ancestry; the latter result showed nominally significant evidence of association. As the association result implicates var321 in increased risk for TS at the population level, we hypothesized that it might be associated with TS in a large cohort of TS patients drawn from two populations. We found no evidence for association of var321 with TS in either population, nor did we observe any occurrences of varCDfs in the Ashkenazi controls with no history of tics or obsessive behaviors.

**RESULTS**

We genotyped the variants varCDfs and var321 in 822 individuals from independently ascertained nuclear families. TS families included 193 TS patients and 322 parents from the Ashkenazi Jewish (AJ) population, and 114 TS patients and 193 parents from the Central Valley of Costa Rica (CVCR). Genotyping data are presented in Table 1. We did not find either variant in the CVCR samples. We also were unable to observe any occurrences of varCDfs in the Ashkenazi cohort. We did, however, identify eight instances of the non-coding variant var321 in heterozygous form: in five unrelated, phenotype-unknown parents of TS probands and in three affected children. All var321 genotypes were confirmed by sequencing, and parentage was verified using microsatellite markers. The transmission patterns of the var321 allele are presented in Figure 1. Whereas three of the var321-carrying parents transmitted var321 to their TS-affected child, two var321-carrying parents did not transmit the allele. These results indicate that the variant does not segregate with disease in our samples.

Our observation of var321 on the non-transmitted parental Ashkenazi chromosomes prompted us to look for the variant in a larger matched Ashkenazi cohort unrelated to TS (Repository for DNA for Ashkenazi Jewish Controls http://www.hopkinsmedicine.org/epigen/). We found var321 in heterozygous form in one individual from a cohort of 256 Ashkenazi controls with no history of tics or obsessive behaviors. Failure of var321 to segregate with disease in two out of five families argues against var321 being a major TS-susceptibility allele. Further, the observation of var321 in Ashkenazi families and controls suggests that var321 has a higher prevalence in the Ashkenazim than in the CVCR, and possibly the European control population assessed by Abelson et al. Although the frequency of var321 in the Ashkenazi control population is not significantly different from European controls ($P = 0.1071$, Fisher’s Exact Test), the probability of not observing any instance of a variant with a frequency of 1/512 (the observed frequency in Ashkenazi control chromosomes) in the 4296 European control chromosomes of Abelson et al. is small ($P = 0.0002251$), suggesting that the Ashkenazi control group does indeed differ from the controls assessed by Abelson et al.

**DISCUSSION**

The overriding issue in conducting association analyses of rare variants is that it is difficult to obtain enough statistical power to unequivocally interpret their results. In our study, the sample of several hundred individuals does not have
sufficient power for our results to be considered as disproving the association described by Abelson et al. Nonetheless, our data suggest a possible alternative explanation that is important to consider: var321 may be an example of a rare variant that is overrepresented (although still infrequent) in the Ashkenazi population. We observed six independent occurrences of var321 in the Ashkenazi population, and found that this variant does not segregate with TS in the families in which it is observed. Our finding that this variant occurs in unaffected Ashkenazi individuals, together with its lack of occurrence in both the large European-derived control sample of Abelson et al. and in the CVCR population, suggests that var321 may be a rare Ashkenazi population variant. The recent observation of one heterozygous var321 in a healthy European-derived control (7) further puts into doubt the association of the variant with TS. The functional studies of SLITRK1 carried out by Abelson et al. do not independently strengthen the evidence for SLITRK1 as a candidate susceptibility gene for TS. Indeed, the preponderance of data indicates that strong statistical evidence for a disease locus is a prerequisite for introducing functional evidence regarding the possible means by which a given gene may play a role in the pathophysiology of a particular disease (8). In the case of SLITRK1, the previous association results for var321 reported by Abelson et al. provided the sole statistical evidence underlying the proposed implication of SLITRK1 in TS etiology.

Although our results do not provide a definitive answer regarding the validity of the association proposed by Abelson et al., they suggest that these results may reflect hidden stratification between cases and controls. As with other aspects of interpreting association results for rare variants, available methods may not have sufficient power to permit clear conclusions regarding this possibility. Furthermore, family-based designs for association studies—which provide the most reliable guard against spurious associations that are due to such stratification—have very little power to detect association for rare variants.

The likelihood that rare variants may be limited to a particular ethnic population underscores the particular importance of ethnically matching case and control samples in association analysis of rare variants. Because there is a higher likelihood of a substantial gradient in allele frequencies between different populations of European origin for rare variants than for common variants, matching cases and controls for ethnic population is more difficult for rare variants. This is particularly true in countries such as the USA in which there is both considerable ethnic admixture and regional variability in the frequency of particular ethnic backgrounds. For example, a rare variant that is overrepresented in the Ashkenazi population is much more likely to be present in samples drawn from areas with current or historically high Jewish populations (e.g. the New York Metropolitan Area) than from regions with smaller Jewish populations (e.g. the rural Midwest). One means of partially allaying this problem is to ascertain cases and controls from within the same geographical area. Such geographical matching may be useful even if the subjects do not identify themselves as members of particular ethnic groups. Ancestral self-identification, which has historically been used to match cases and controls, may be a relatively unreliable source of information. For example, a recent study assessed the reliability of self-reported ancestry among siblings residing in Minnesota and surrounding states and found that less than half of siblings agreed on the countries of origin of both parents (9).

Given the unreliability of self-reported ethnicity as a guard against stratification, several methods have been developed to detect stratification using genotype data. Failure to detect stratification using these approaches should be interpreted cautiously. For example, screening for the presence of hidden population stratification (e.g. using random unlinked markers) may not detect subtle differences in the genetic composition of case and control groups, which can contribute substantially to spurious association results (10). It may be possible to improve the power of such analyses by obtaining more extensive self-report genealogies from cases and controls (11). Genomic control approaches (12) are becoming widely used, but may not be sensitive to even moderate levels of stratification (13). Correction factors may be applied that enhance the utility of this approach, in that—even when there is no direct evidence of stratification—the null hypothesis is rejected at more appropriate levels (13). Typing cases and controls for ancestry-informative markers (AIM) (14) may be useful for interpreting associations for rare variants that are overrepresented within a particular ethnic population. However, for this approach to be generally useful, it will be necessary to employ AIMs that are in linkage disequilibrium with the markers being tested for association, requiring the identification of such markers at high resolution across the genome.

Cases and control samples may be mismatched for reasons other than population stratification. For example, differential processing of case and control DNAs (which are often collected separately and even by different groups of investigators) may generate occasional genotyping artifacts that may be misinterpreted as rare variants. It has been proposed that statistical methods such as genomic control analysis may be useful for detecting such occurrences (15).

While our data highlight the special problem of ethnically matching cases and controls for association studies of rare variants, they also add to the growing literature indicating that interpreting association tests between rare variants and complex traits is a general problem for the field. For example, a recent report implicated a rare 21 bp deletion in MEF2A in susceptibility to coronary artery disease (CAD) (16). Subsequently Weng et al. (17) identified this MEF2A deletion in six elderly control individuals from three families who showed no evidence of CAD, suggesting that the initial report might reflect a chance co-occurrence. In the cases of both SLITRK1 and MEF2A, the variants in question appear to be rare polymorphisms that do not consistently segregate with their respective diseases. Clearly, statistically significant mutations, multiple mutations that are functional and co-segregate with disease, de novo mutation, and/or model systems are required to prove a link between variation in these genes and disease (3).

It is not yet clear what strategy will be best for assessing the significance of association to rare variants. Systematically
screening candidate genes in both cases and controls and then jointly analyzing the complete set of variants to test for association in aggregate has already been proven feasible as a means to obtain adequate statistical power to analyze associations for rare variants in genes of at least moderate effect on complex traits (e.g. those contributing to different serum levels of lipoproteins) (2). This strategy has the advantage that it does not place excessive weight on the findings for any single variant. Testing such aggregates of rare variants in cases and controls requires systematic variant discovery in the two groups for all candidate genes being considered in relation to a disease. In addition, it is not certain that strategies involving the joint analysis of several variants together will guard against hidden stratification; while it seems unlikely that a similar pattern of stratification has the advantage that it does not place excessive weight on the need for the field to develop a consensus on appropriate approaches for design of association analysis of rare variants.

MATERIALS AND METHODS

Samples

The research protocol was approved by the UCLA, UCSD, UCSF, Johns Hopkins and Hospital Nacional de Niños (Costa Rica) IRBs. TS probands were recruited through clinical referral, the Tourette Syndrome Association, newspaper and television advertisements describing key features of the syndrome, schools and family referral. All probands met strict diagnostic and genealogic criteria. The study included only patients with diagnoses of definite TS/definite TS, definite TS/TS history, or probable TS/definite TS by DSM-IV and TS Collaborative Study Group (TSCSG) criteria. Parents were not routinely clinically assessed, and are phenotype unknown unless otherwise indicated. Ashkenazi probands have at least three of four grandparents of Ashkenazi descent, and CVCR probands have at least five of eight great-grandparents from the CVCR. All individuals found to have the var321 allele had all four grandparents of Ashkenazi descent, and CVCR probands have at least five of eight great-grandparents from the CVCR. All individuals found to have the var321 allele had all four grandparents of Ashkenazi descent. AJ controls were recruited for the study of complex diseases in Jews, and were ascertained from multiple sources including advertisement and presentations. Individuals selected for this study had no history of neuropsychiatric traits (self-reported from medical and psychiatric history questionnaires, which included questions about tics and obsessions and compulsions), had all four grandparents of Ashkenazi descent, and were >40 years old at the time of collection.

Genotyping

Genotype data were collected via TaqMan Assay-by-Design on a 7900HT PCR system (Applied Biosystems) using the following primer sequences: SLITRK1-V321F, 5'-CATGATAACTGAGTGCATTTTGC-3'; SLITRK1-V321R, 5'-ACATTTTTTCAGTTAATCGTCTG GAAC-3'; SLITRK1-V321V1, 5'-VIC-ATTATTTGTGATCGTG AGGC A-3'; SLITRK1-V321M1, 5'-FAM-ATTATTTGTGATCGTG AGGC A-3'; SLITRK1-CDFSV1, 5'-VIC-AC ACTTCAAGACCTTTTGG-3'; SLITRK1-CDFSM1, 5'-F AM-CTT TCAAGACCTTTTGG-3'. Sequencing to verify var321 genotypes was accomplished using M13 forward and reverse primers and the following SLITRK1-specific amplification primers: 5'-TGTA AACAGCAGCCGATTGTGCAAG AGGCATGAATG-3'; 5'-AGGAAACAGCATTGACCATAGC TAAAGCCTAGAGTGAC-3'.

ACKNOWLEDGEMENTS

The authors thank the patients and families who participated in the study, the Tourette Syndrome Association, individuals who provided DNA to the Repository for Ashkenazi Jewish Controls, and the Tourette Syndrome Association International Consortium on Genetics (TSAICG). Support for this work was provided by grants from the NINDS, R01-NS037484 (N.B.F.) and R01-NS040024 (TSAICG). D.K.K. is supported by training grants from the Tourette Syndrome Association Inter-continental Network of Centers for Research on Genetics (TSAICG). Support for this work was provided by grants from the NINDS, R01-NS037484 (N.B.F.) and R01-NS040024 (TSAICG). They also thank Peter Heutink and Sue Service for comments on the manuscript and members of the TSAICG for discussions about the data.

Conflict of Interest statement. No conflicts declared.

REFERENCES


