Neuropsychopharmacology (2012) 37, 557–566
© 2012 American College of Neuropsychopharmacology. All rights reserved 0893-133X/12
www.neuropsychopharmacology.org

# Genome-Wide Association Study of Alcohol Dependence Implicates *KIAA0040* on Chromosome 1q

**Lingjun Zuo*,[1], Joel Gelernter[1,2,3], Clarence K Zhang[4], Hongyu Zhao[4], Lingeng Lu[4], Henry R Kranzler[5], Robert T Malison[1,6], Chiang-Shan R Li[1], Fei Wang[1], Xiang-Yang Zhang[7], Hong-Wen Deng[8], John H Krystal[1,3,9], Fengyu Zhang[10] and Xingguang Luo*,[1]**

*[1]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA; [2]Departments of Genetics and Neurobiology, Yale University School of Medicine, New Haven, CT, USA; [3]Alcohol Research Center, VA Connecticut Healthcare System, West Haven, CT, USA; [4]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, USA; [5]Department of Psychiatry, University of Pennsylvania and MIRECC, Philadelphia VAMC, Philadelphia, PA, USA; [6]Clinical Neuroscience Research Unit, Connecticut Mental Health Center, New Haven, CT, USA; [7]Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, USA; [8]Department of Biostatistics, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, USA; [9]Yale-New Haven Hospital, New Haven, CT, USA; [10]Gene, Cognition and Psychosis Program, National Institute of Mental Health, National Institutes of Heath, Bethesda, MD, USA*

Previous studies using SAGE (the Study of Addiction: Genetics and Environment) and COGA (the Collaborative Study on the Genetics of Alcoholism) genome-wide association study (GWAS) data sets reported several risk loci for alcohol dependence (AD), which have not yet been well replicated independently or confirmed by functional studies. We combined these two data sets, now publicly available, to increase the study power, in order to identify replicable, functional, and significant risk regions for AD. A total of 4116 subjects (1409 European-American (EA) cases with AD, 1518 EA controls, 681 African-American (AA) cases, and 508 AA controls) underwent association analysis. An additional 443 subjects underwent expression quantitative trait locus (eQTL) analysis. Genome-wide association analysis was performed in EAs to identify significant risk genes. All available markers in the genome-wide significant risk genes were tested in AAs for associations with AD, and in six HapMap populations and two European samples for associations with gene expression levels. We identified a unique genome-wide significant gene—*KIAA0040*—that was enriched with many replicable risk SNPs for AD, all of which had significant *cis*-acting regulatory effects. The distributions of −log(*p*) values for SNP-disease and SNP-expression associations for all markers in the *TNN–KIAA0040* region were consistent across EAs, AAs, and five HapMap populations ($0.369 \leqslant r \leqslant 0.824$; $2.8 \times 10^{-9} \leqslant p \leqslant 0.032$). The most significant SNPs in these populations were in high LD, concentrating in *KIAA0040*. Finally, expression of *KIAA0040* was significantly ($1.2 \times 10^{-11} \leqslant p \leqslant 1.5 \times 10^{-6}$) associated with the expression of numerous genes in the neurotransmitter systems or metabolic pathways previously associated with AD. We concluded that *KIAA0040* might harbor a causal variant for AD and thus might directly contribute to risk for this disorder. *KIAA0040* might also contribute to the risk of AD via neurotransmitter systems or metabolic pathways that have previously been implicated in the pathophysiology of AD. Alternatively, *KIAA0040* might regulate the risk via some interactions with flanking genes *TNN* and *TNR*. *TNN* is involved in neurite outgrowth and cell migration in hippocampal explants, and *TNR* is an extracellular matrix protein expressed primarily in the central nervous system.

*Neuropsychopharmacology* (2012) **37,** 557–566; doi:10.1038/npp.2011.229; published online 28 September 2011

**Keywords:** risk region; alcohol dependence; *cis*-eQTL; GWAS

## INTRODUCTION

Alcohol dependence (AD) is a complex disorder characterized by psychological and physiological dependence on ethanol. The 12-month prevalence of AD in the United States is 3.81% (Grant *et al*, 2004). Family, twin, and adoption studies have demonstrated that genetic factors constitute a significant component of the risk for AD. Candidate gene studies have shown that a large number of risk loci exist for AD in the dopaminergic, serotonergic, GABAergic, cholinergic, opioidergic, and endocannabinoidergic systems, as well as in the ethanol metabolic pathway. Several genome-wide association studies (GWASs) have reported additional risk loci for alcoholism (Treutlein *et al*, 2009; Bierut *et al*, 2010; Edenberg *et al*, 2010; Heath *et al*, 2011). The first GWAS in German males reported that

*Correspondence: Dr L Zuo or Dr X Luo, Department of Psychiatry, Yale University School of Medicine, West Haven, CT 06516, USA, Tel: + 1 203 932 5711 ext 3590, Fax: + 1 203 937 4741, E-mail: Xingguang.Luo@yale.edu or Lingjun.Zuo@yale.edu

Received 26 July 2011; revised 27 August 2011; accepted 29 August 2011

15 top-ranked SNPs (in *PECR*, *ADH1C*, *CAST*, *ERAP1*, *PPP2R2B*, *ESR1*, *GATA4*, *CCDC41*, and *CDH13*) ($5.6 \times 10^{-6} \leqslant p \leqslant 2.2 \times 10^{-3}$) in a discovery sample were replicated in a follow-up data set; two in *PECR* (2q35) reached genome-wide significance in meta-analysis ($\alpha = 5 \times 10^{-8}$) (Treutlein *et al*, 2009). But the top-ranked SNPs ($p < 10^{-4}$) in this German discovery sample were not replicated by a second GWAS for AD (Bierut *et al*, 2010), which reported a different set of 15 top-ranked SNPs (in *PKNOX2*, *CC2D2B*, *NOMO2*, *COL8A1*, *NXPH2*, *E2F8*, *FAM44B*, *SH3BP5*, *GRM5*, *ZNF285A*, and *TPK1*) in a combined European-American (EA) and African-American (AA) sample from SAGE (the Study of Addiction: Genetics and Environment), all of which were nominally associated with alcoholism ($1.9 \times 10^{-7} \leqslant p \leqslant 9.8 \times 10^{-6}$). However, these SNPs were neither genome-wide significant nor were they replicated in a family sample from COGA (the Collaborative Study on the Genetics of Alcoholism) or the German case–control sample. Furthermore, the top-ranked SNPs ($p < 10^{-6}$) in EAs (or AAs) were not replicated in AAs (or EAs) (Bierut *et al*, 2010). The third GWAS found no SNP with genome-wide significance in an EA COGA discovery sample. However, 6 SNPs in *TMEM132C*, *EPHA7*, *OPA3*, *KCNMA1*, *DMRTA2*, and *SPTA1* for AD and 41 SNPs in *SELL*, *SELE*, *LOC91431*, *PPARG*, *CTNN2*, *LEPR*, and *PDE4B* for early-onset AD were replicated in an AA replication sample. Ten SNPs in *CARS*, *OSBPL5*, *NAP1L4*, *BBX*, *SLC9A8*, *OPA3*, *TOX2*, and *CD53* for AD and 16 SNPs in *SLC37A3*, *KCNMA1*, *CDH8*, *ZNF608*, *API5*, *CAT*, and *GRIN2C* for early-onset AD were replicable between the EA case–control discovery sample and an EA family replication sample (Edenberg *et al*, 2010). Most recently, a GWAS in an Australian twin-family sample identified *TMEM108* and *ANKS1A* as possible risk genes for alcohol consumption (Heath *et al*, 2011). Another GWAS meta-analysis in European populations identified *AUTS2* as the risk gene for alcohol consumption (Schumann *et al*, 2011). However, these findings have not yet been replicated in independent samples or confirmed by functional studies, and hence the possibility of a false-positive result cannot yet be excluded.

In the present study, we combined and reanalyzed the SAGE and COGA data sets, and used a new analytic strategy to identify additional risk loci for AD. First, we combined both data sets, hoping to increase the sample sizes and, in turn, the study's statistical power (site effects and sample overlapping were taken into account), thereby enhancing our ability to detect novel risk loci that were missed in previous studies. Second, we differentiated more fully the EAs and AAs in the analysis to increase population homogeneity, and controlled for admixture effects in the association tests. Third, we used the EAs as a discovery sample and the AAs as a replication sample, and different samples with distinct ethnicity to detect expression quantitative trait locus (eQTL) signals, as a confirmation of the variants' functional effects. Although using distinct samples in one study might increase the false-negative rates due to sample heterogeneity, replication in distinct samples makes the false-positive findings less likely. Replicable findings in distinct populations might be more generalizable to other populations, and would be more likely to be causal in nature. Fourth, allele frequencies could be different in distinct populations, or even exist in opposite phases; that is, a common allele in one population may be a rare allele in another population. Thus, distinct populations do not necessarily have the same risk markers associated with disease; alternatively, they could have the same risk markers, but have different phases of alleles in these markers associated with disease. That is, the effect sizes and effect directions of marker–disease associations may be not consistent, or could even be opposite across populations for each individual risk marker, such that meta-analysis may show weaker effects. Such markers are usually treated as nonreplicable and thus are discarded. However, to our hypothesis, when two distinct populations have common causal variants, there could be a risk region in LD with this putative causal variant in both populations, even though there are no individual risk alleles replicable between them. This is because in one population, a set of risk markers are in LD with the causal variant; but in another population, a different set of risk markers adjacent to the first set could be in LD with the causal variant. The risk marker sets in a causal region are different between populations, because they are not causal variants *per se*. Such a risk region may have a significant correlation between the distributions of $-\log(p)$ values of all markers across the entire region in different populations. Such regions have usually been missed in previous studies. Additionally, in the present study, the data sets for association studies and for eQTL analysis were different in many absolute statistics of genetic marker numbers, sample sizes, and study power. To study the consistency between them, we can only compare their relative statistics, that is, the distributions of relative significance strengths across whole regions, not individual markers. In a word, the present study aimed to identify replicable, functional, and significant risk regions for AD by increasing the study power and the sample homogeneity.

## MATERIALS AND METHODS

### Subjects

A total of 4316 SAGE (dbGaP study accession phs000092.v1.p1) subjects and 1957 COGA (dbGaP study accession phs000125.v1.p1) subjects were merged into a single data set; 1477 subjects in COGA who overlapped with SAGE were excluded. The demographic data of SAGE and COGA subjects have been presented previously (Bierut *et al*, 2010; Edenberg *et al*, 2010). After data cleaning (see below), 1409 EA cases (37.3% females; 38.3 ± 10.2 years), 1518 EA controls (70.7% females; 39.4 ± 10.4 years), 681 AA cases (37.2% females; 40.3 ± 7.8 years), and 508 AA controls (66.7% females; 39.6 ± 8.6 years) underwent analysis. Affected subjects met lifetime DSM-IV criteria (American Psychiatric Association, 1994) for AD. Controls were defined as individuals who had been exposed to alcohol (and possibly to other drugs) in sufficient amounts for a sufficient time, but had never become addicted to alcohol or other illicit substances (lifetime diagnoses). Additionally, controls were also screened to exclude individuals with major axis I disorders, including schizophrenia, mood disorders, and anxiety disorders. More demographic data are provided in Supplementary Materials and Methods. All subjects were de-identified in this study. All subjects were genotyped on the Illumina Human 1M beadchip.

## Data Cleaning

Subjects with poor genotypic data and questionable diagnostic information, subjects with allele discordance, duplicated IDs, potential sample misidentification, sample relatedness, sample misspecification, gender anomalies, chromosome anomalies (such as aneuploidy and mosaic cell populations), missing race, non-EA and non-AA ethnicity, population group outliers, subjects with a mismatch between self-identified and genetically inferred ethnicity, and subjects with a missing genotype call rate $\geqslant$2% across all SNPs were excluded step by step (Supplementary Table S1). Furthermore, SNPs with allele frequency difference in controls $>$2% between SAGE and COGA, SNPs with missing rate difference $>$2% between SAGE and COGA, and SNPs with allele discordance, chromosomal anomalies, or batch effect were excluded. We then filtered out the SNPs on all chromosomes with an overall missing genotype call rate $\geqslant$2%, the monomorphic SNPs, and the SNPs with minor allele frequencies (MAFs) $<$0.01 in either EAs or AAs. The SNPs that deviated from Hardy–Weinberg equilibrium (HWE; $p<10^{-4}$) within EA or AA controls were also excluded. This selection process yielded 805 814 SNPs in EAs and 895 714 SNPs in AAs. Details are provided in Supplementary Materials and Methods.

## Association Analysis

(a) Genome-wide association tests in the EA discovery sample: The allele frequencies were compared between cases and controls in EAs using genome-wide logistic regression analysis implemented in the program PLINK (Purcell *et al*, 2007). Diagnosis served as the dependent variable, alleles served as the independent variables, and ancestry proportions, sex, and age served as covariates. The *p*-values derived from these analyses are illustrated in Supplementary Figure S1 and the top-ranked risk SNPs ($p<10^{-5}$) are listed in Table 1.

(b) Association tests in the AA replication sample: Associations between the top-ranked risk SNPs in EAs were tested using logistic regression analysis in AAs. Additionally, associations for all available SNPs in the genome-wide significant risk genes identified in EAs were also tested in AAs (Figure 1). Meta-analysis was performed to derive the combined *p*-values for EAs and AAs.

(c) Controlling for admixture effects: The ancestry proportions for each individual were estimated by integrating the ancestry information content of 3172 completely independent ancestry-informative markers (AIMs) using STRUCTURE (Pritchard *et al*, 2000). These AIMs were extracted from 1 million of markers by LD pruning (Purcell *et al*, 2007) (see details in Supplementary Materials and Methods). To control for the admixture effects on association analysis, ancestry proportions were included as covariates in the association analysis.

**Table 1** Top-Ranked SNPs in EAs with $P<10^{-5}$ for Associations

| Band | SNP | Genes | *P*-value | FDR | OR |
|---|---|---|---|---|---|
| 1q24–q25 | rs1057239[a] | *KIAA0040* | $2.8\times10^{-7}$ | 0.038 | 1.32 |
| 1q24–q25 | rs1894709[a] | *KIAA0040* | $3.9\times10^{-7}$ | 0.039 | 1.31 |
| 1q23–q24 | rs6701037[a] | *KIAA0040* | $1.3\times10^{-6}$ | 0.043 | 1.30 |
| 1q24–q25 | rs6425323[a] | *KIAA0040* | $1.9\times10^{-6}$ | 0.043 | 1.29 |
| 1q24–q25 | rs1057302[a] | *KIAA0040* | $2.5\times10^{-6}$ | 0.044 | 1.29 |
| 1p35.1 | rs4949400[a] | *SERINC2* | $2.3\times10^{-7}$ | 0.043 | 1.32 |
| 1p35.1 | rs4949402[a] | *SERINC2* | $2.6\times10^{-7}$ | 0.043 | 1.32 |
| 1p35.1 | rs1039630[a] | *SERINC2* | $2.6\times10^{-7}$ | 0.043 | 1.32 |
| 1p35.1 | rs2275435[a] | *SERINC2* | $3.0\times10^{-7}$ | 0.044 | 1.31 |
| 1p35.1 | rs4478858[a] | *SERINC2* | $4.4\times10^{-7}$ | 0.044 | 1.31 |
| 1p35.1 | rs10914386 | *SERINC2* | $9.0\times10^{-6}$ | >0.05 | 1.30 |
| 5q12.1 | rs7445832[a] | *HTR1A* | $2.8\times10^{-7}$ | 0.041 | 1.38 |
| 1p34.3 | rs11583322 | *STK40* | $4.0\times10^{-7}$ | >0.05 | 0.76 |
| 5q23.3 | rs257906 | *SLC27A6* | $5.0\times10^{-7}$ | >0.05 | 1.43 |
| 5q23.3 | rs10478829 | *SLC27A6* | $5.3\times10^{-7}$ | >0.05 | 1.43 |
| 10q25.3 | rs4751971 | *PNLIPRP3* | $2.4\times10^{-6}$ | >0.05 | 3.60 |
| 13q11 | rs1867248 | *TUBA3C* | $2.6\times10^{-6}$ | >0.05 | 1.33 |
| 12q21.3 | rs882968 | *ATP2B1* | $3.6\times10^{-6}$ | >0.05 | 1.29 |
| 5q11.2–q13 | rs4700575 | *HTR1A* | $3.8\times10^{-6}$ | >0.05 | 1.35 |
| 5q11.2–q13 | rs2169520 | *HTR1A* | $4.0\times10^{-6}$ | >0.05 | 1.35 |
| 13q12 | rs7983722 | *ATP8A2* | $4.1\times10^{-6}$ | >0.05 | 6.48 |
| 12q24.1 | rs4964684 | *CMKLR1* | $5.1\times10^{-6}$ | >0.05 | 4.18 |
| 12q23.3 | rs1896086 | *KIAA0789* | $5.4\times10^{-6}$ | >0.05 | 4.37 |
| 12q23.3 | rs7971309 | *KIAA0789* | $5.4\times10^{-6}$ | >0.05 | 4.37 |
| 3p13 | rs6802792 | *RYBP* | $5.5\times10^{-6}$ | >0.05 | 9.80 |
| 7q31 | rs17142876 | *KCND2* | $5.6\times10^{-6}$ | >0.05 | 1.51 |
| 7q31 | rs728115 | *KCND2* | $5.6\times10^{-6}$ | >0.05 | 1.49 |
| 8q11–q12 | rs9298318 | *SNTG1* | $6.9\times10^{-6}$ | >0.05 | 6.27 |
| 8q13.3 | rs12549296 | *EYA1* | $8.8\times10^{-6}$ | >0.05 | 1.81 |
| 11q24.2 | rs750338 | *PKNOX2* | $9.5\times10^{-6}$ | >0.05 | 1.32 |
| 2q22.1 | rs1869324 | *THSD7B* | $9.8\times10^{-6}$ | >0.05 | 1.44 |
| 5q14 | rs6888626 | *COX7C* | $9.9\times10^{-6}$ | >0.05 | 6.66 |
| 12q24.33 | rs2218917 | *TMEM132D* | $1.0\times10^{-5}$ | >0.05 | 9.04 |

Abbreviations: FDR, genome-wide false discovery rate; OR, odds ratio.
[a]Indicates FDR$<$0.05.

## Functional Analysis (*Cis*-Acting Genetic Regulation of Expression Analysis)

(a) *Cis*-acting expression of QTL (*Cis*-eQTL) analysis on the risk SNPs in lymphoblastoid cell lines: To examine relationships between all risk SNPs (Table 2) and local mRNA expression levels, we used expression data of 14 925 transcripts (14 072 genes) in 270 unrelated HapMap individuals from six populations (CEU-Children, CEU-Parent, CHB, JPT, YRI-Children, and YRI-Parent) (Stranger *et al*, 2005). Differences in the distribution of mRNA expression levels between SNP genotypes were compared using a Wilcoxon-type trend test. The *p*-values of $<$0.05 are listed in Table 2. Effects of SNPs 1 Mb around the association peak SNP (rs1057239) are illustrated in Figure 1d–i.

(b) *Cis*-eQTL analysis on the risk SNPs in other two primary human cells: To examine whether the risk SNPs influence the local exon- or transcript-level expression changes, we also tested the associations between the genotypes of these risk SNPs and the expression levels of exons and transcripts of local genes in the other two European samples (Table 2 and Figure 1j and k).

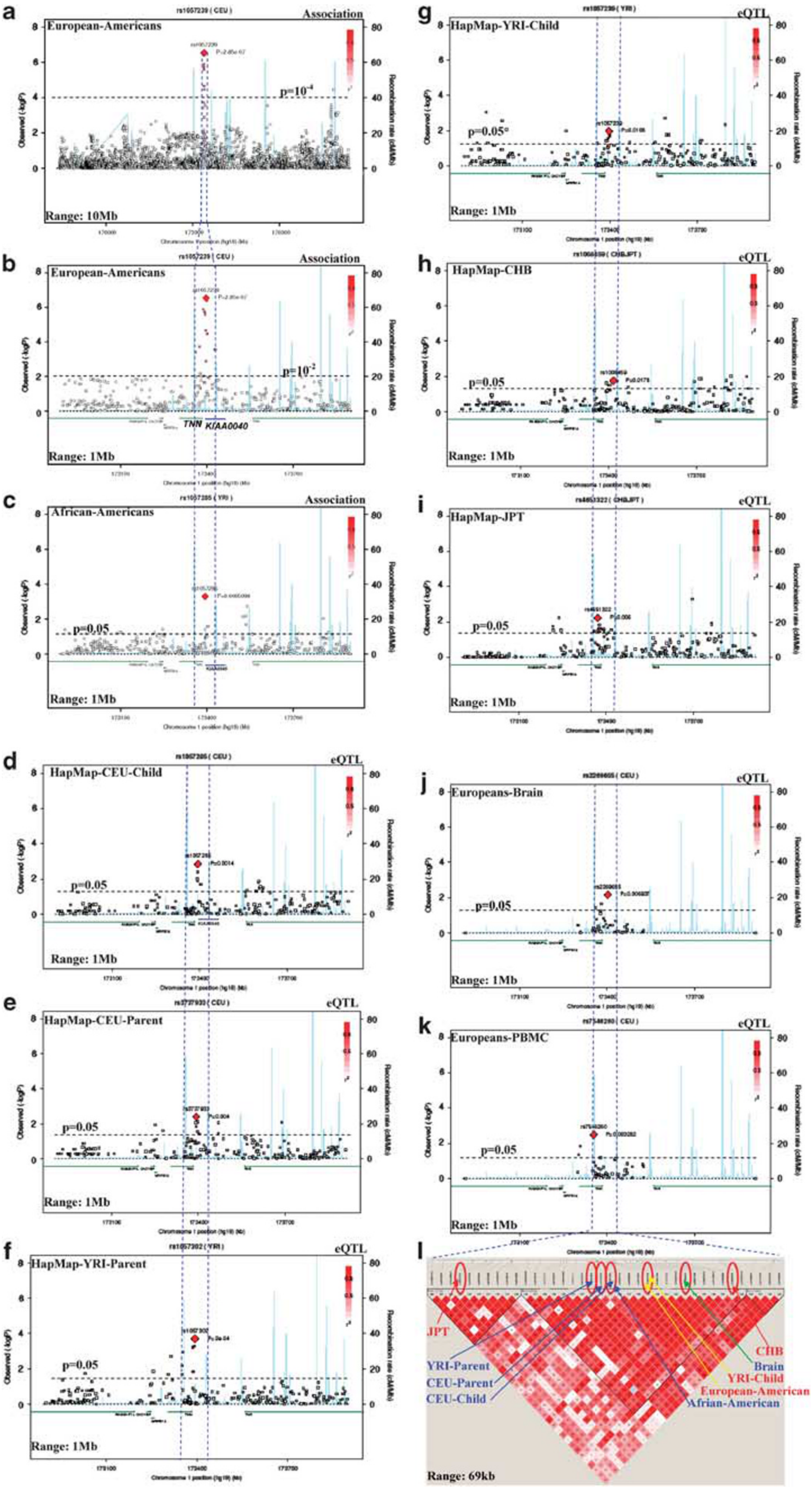

**Figure 1** Regional association and eQTL plots around *TNN–KIAA0040* region. Left y-axis corresponds to −log(p) value; right y-axis corresponds to recombination rates; quantitative color gradient corresponds to $r^2$; red squares represent peak SNPs. (a) Regional association plot in EAs for a 10 Mb region around the peak association SNP (rs1057239); (b, c) regional association plots in EAs and AAs for a 1 Mb region around the peak association SNPs; (d–k) regional eQTL plots in HapMap and European populations for a 1 MB region around the peak functional SNPs; and (l) LD map for *TNN–KIAA0040* region.

**Table 2** *P*-Values for Associations and *Cis*-Acting Regulatory Effects of the SNPs in *TNN–KIAA0040* Region

| SNP | A1 | EA | | AA | | *P*-values for eQTL in six HapMap populations | | | | | | *P*-values for eQTL in Europeans[a] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | OR | *P*-value | OR | *P*-value | CEU_children | CEU_parent | CHB | JPT | YRI_children | YRI_parent | Transcript (B) | Exon 1 (P) | Exon 4 (P) |
| rs12094153 | A | 1.210 | $3.9 \times 10^{-4}$ | — | — | — | — | — | — | — | — | — | $2.8 \times 10^{-3}$ | — |
| rs4651322 | T | 0.889 | 0.034 | — | — | — | — | — | $6.0 \times 10^{-3}$ | — | — | — | — | — |
| rs2157588 | A | 1.219 | $1.9 \times 10^{-4}$ | — | — | — | — | — | — | — | — | — | — | — |
| rs12563833 | A | 0.893 | 0.047 | — | — | — | — | — | 0.023 | — | — | — | — | — |
| rs1018829 | T | 0.839 | $6.8 \times 10^{-3}$ | — | — | — | — | — | 0.035 | — | — | — | — | — |
| rs10489328 | A | 0.848 | 0.017 | — | — | — | — | — | — | — | — | 0.023 | — | — |
| rs2018318 | T | 0.836 | $9.9 \times 10^{-3}$ | — | — | — | — | — | — | — | — | — | — | — |
| rs6701037 | C | 1.295 | $1.3 \times 10^{-6}$ | — | — | — | — | — | 0.025 | 0.037 | $6.0 \times 10^{-4}$ | — | $9.4 \times 10^{-4}$ | — |
| rs10912899 | A | 0.871 | $9.4 \times 10^{-3}$ | — | — | — | — | — | — | — | 0.014 | — | — | $3.2 \times 10^{-4}$ |
| rs4650707 | A | 0.741 | 0.024 | — | — | — | — | — | — | — | — | — | — | — |
| rs6425323 | T | 1.291 | $1.9 \times 10^{-6}$ | — | — | 0.010 | $8.9 \times 10^{-3}$ | — | 0.025 | 0.038 | $5.0 \times 10^{-4}$ | — | — | — |
| rs1057302 | C | 1.286 | $2.5 \times 10^{-6}$ | — | — | 0.013 | $4.8 \times 10^{-3}$ | — | 0.025 | — | $2.0 \times 10^{-4}$ | — | — | — |
| rs3737933 | C | 0.850 | 0.023 | 1.757 | 0.040 | $4.0 \times 10^{-3}$ | $4.0 \times 10^{-3}$ | 0.024 | 0.035 | — | — | — | — | — |
| rs1057285 | G | — | — | 1.412 | $5.1 \times 10^{-4}$ | $1.4 \times 10^{-3}$ | 0.014 | — | — | — | — | — | — | — |
| rs4650716 | C | 1.249 | $2.2 \times 10^{-5}$ | — | — | — | — | — | — | 0.025 | — | — | — | — |
| rs2269650 | A | 0.842 | $1.4 \times 10^{-3}$ | — | — | — | — | — | — | — | — | — | — | $8.6 \times 10^{-5}$ |
| rs2072035 | A | 0.841 | $1.3 \times 10^{-3}$ | — | — | — | — | — | — | — | — | — | — | — |
| rs1057239 | T | 1.318 | $2.8 \times 10^{-7}$ | — | — | — | 0.028 | — | — | 0.011 | — | — | — | — |
| rs1894709 | A | 1.313 | $3.9 \times 10^{-7}$ | — | — | — | $8.0 \times 10^{-3}$ | — | — | 0.016 | — | — | — | — |
| rs10489326 | T | 1.242 | $3.4 \times 10^{-5}$ | — | — | — | — | — | — | 0.021 | — | — | — | — |
| rs2269655 | T | — | — | 1.875 | 0.026 | 0.020 | 0.039 | 0.031 | — | — | — | $6.9 \times 10^{-3}$ | — | — |
| rs2861158 | A | 0.863 | 0.011 | — | — | — | — | — | — | — | — | — | — | — |
| rs16847872 | G | — | — | 1.872 | 0.026 | 0.020 | 0.039 | 0.031 | — | — | — | — | — | — |
| rs12136973 | C | 0.865 | 0.012 | — | — | — | — | — | 0.040 | — | — | — | — | $2.5 \times 10^{-3}$ |
| rs1008459 | G | 0.881 | 0.029 | — | — | — | — | 0.018 | 0.028 | — | — | — | — | $1.1 \times 10^{-3}$ |
| rs2272785 | A | 0.879 | 0.025 | — | — | — | — | 0.018 | 0.044 | — | — | — | — | — |
| rs2272784 | C | 0.823 | $2.8 \times 10^{-4}$ | — | — | — | — | — | — | — | — | — | — | $7.7 \times 10^{-6}$ |
| rs3753555 | C | 0.823 | $2.8 \times 10^{-4}$ | — | — | — | — | — | — | — | — | — | — | $7.7 \times 10^{-6}$ |

Abbreviations: eQTL, expression quantitative trait locus analysis.

[a] Exon-level expression changes in brain (B) and PBMC (P) tissues were corrected for 4 exons and 2 tissues ($\alpha = 0.0063$).

'—', $p > 0.05$ or not available.

All markers are in HWE.

Expression data in 93 autopsy-collected frontal cortical brain tissue samples with no defined neuropsychiatric condition and 80 peripheral blood mononuclear cell (PBMC) samples collected from living healthy donors obtained from a study (Heinzen *et al*, 2008) at Duke University (Duke samples) were evaluated. Each of these associations was analyzed using a linear regression model by correcting for age, sex, source of tissues, and principle component scores.

## Correction for Multiple Testing on Association and *Cis*-eQTL Analysis

To mitigate false-positive rates, genome-wide associations in the discovery stage need to be corrected for multiple testing. Apparently, Bonferroni correction ($\alpha = 5 \times 10^{-8}$) was overly conservative because it treated all 1 million markers in the genome as independent ones (which is impossible). Alternatively, a WTCCC-defined $\alpha$ ($= 5 \times 10^{-7}$) might be more appropriate to the present study (The Wellcome Trust Case Control Consortium, 2007). As complements to this correction, we also corrected the discovery findings by genome-wide false discovery rate (FDR; Benjamini and Hochberg, 1995) and replicated and confirmed the discovery findings by replication and confirmation designs. Only when an association survived WTCCC-defined genome-wide correction ($p < 5 \times 10^{-7}$) together with FDR < 0.05, and was replicated in an independent sample and confirmed by functional studies, should it be taken as 'significant'. In the present study, we used multiple samples to replicate and confirm the discovery findings, which significantly reduced the chance of false-positive findings (ie, FDR). First, we used AAs, the population most genetically distinct from EAs in the United States, to replicate the association analysis and make the replicable findings more generalizable to other populations. Second, we aimed to detect replicable regions, not individual markers, to avoid missing risk regions because of the population specificity of allele frequencies introduced above. Many risk markers, rather than a single marker, were detected in the risk regions, which reduced the chance of false-positive association findings. Third, functional analysis as confirmation of association analysis further reduced the chance of false-positive findings. Additionally, functional analysis in multiple different populations, which differed from the populations for association analysis, made the findings more generalizable too. Fourth, $-\log(P)$ value distributions across the discovery sample and the replication and confirmation samples were compared for the similarity using Pearson correlation analysis. The consistency between them would significantly reduce the chance of false-positive findings. Therefore, $\alpha$ could be set at 0.05 for the findings in replication and confirmation samples if they replicated or confirmed the discovery findings (except for exon-level *cis*-eQTL findings that were corrected for the number of exons and the types of tissues; ie, $\alpha$ was set at 0.0063 for *KIAA0040*; see Table 2).

## Functional Analysis (*Trans*-Acting Genetic Regulation of Expression Analysis)

(a) Transcriptome-wide *trans*-eQTL analysis on the risk SNPs: To examine whether the risk SNPs regulated other transcript expression, we tested the associations between the genotypes of these risk SNPs and the transcript expression levels across the transcriptome. The transcriptome-wide expression levels in two human primary cells (brain and PBMC) in the Duke samples (Heinzen *et al*, 2008) were assessed using Affymetrix Human ST 1.0 exon arrays. Associations between genotypes and transcriptome-wide expression levels were analyzed using linear regression implemented in PLINK (Purcell *et al*, 2007) by incorporating all covariates. A total of 2 047 023 transcript expression data records in the brain set, 1 760 880 transcript expression data records in the PBMC set, and 28 risk SNPs were tested, and hence $\alpha$ was set at $8.7 \times 10^{-10}$ for the brain set and $1.0 \times 10^{-9}$ for the PBMC set, respectively.

(b) Genome-wide *trans*-eQTL analysis of *KIAA0040* transcript expression: To examine what polymorphisms across genome regulated the transcript expression of *KIAA0040*, we scanned the whole genome and tested the associations between the transcript expression level of *KIAA0040* and the genotypes across whole genome. The same Duke samples including both tissues as described above were tested (Heinzen *et al*, 2008). Genome-wide genotyping was performed using Illumina Human Hap550K chip. Strict data cleaning was performed before association analysis using previously published methods (Fellay *et al*, 2007). Associations between *KIAA0040* transcript expression level and genotypes across genome were analyzed using linear regression implemented in PLINK, by incorporating all covariates. A total of 571 738 SNPs and 2 tissues were tested, and hence $\alpha$ was set at $4.4 \times 10^{-8}$.

(c) Transcriptome-wide expression correlation analysis: The expression of 14 925 transcripts was examined in Duke samples (Heinzen *et al*, 2008). Correlations between expression of *KIAA0040* transcript and expression of other genes across transcriptome were tested in the brain and PBMC, respectively. The $\alpha$ was set at $1.7 \times 10^{-6}$ ($= 0.05/(14\,925 \times 2)$ in which '2' is the types of tissues tested).

## Functional Analysis (RNA Secondary Structure Analysis)

Each unique DNA sequence across a gene, whether common, rare, or with a unique mutation, could have a different consequent RNA secondary structure. Alteration of RNA secondary structure could influence the efficiency of splicing, translation, and/or binding of regulatory factors. These influences could affect disease susceptibility. Because a test of this hypothesis is beyond the scope of the present report, we used the program MFOLD (Zuker, 2003) to predict an alteration in RNA secondary structures. The upstream and downstream sequences (800 bp) around a variant were retrieved from NCBI dbSNP based on the SNP accession number. The RNA secondary structures of the retrieved sequences with either the common or variant allele were constructed by MFOLD. All parameters were set in default for the most stable secondary structure folding prediction. A $\Delta$G value for each structure corresponding to each allele was derived (Supplementary Figure S2). The $\Delta$G is a metric of stickiness when constructing the

RNA secondary structure, where stickiness represents how thermodynamically stable a structure may be. The larger the absolute value of a negative ΔG is, the more stable a structure may be; conversely, the larger the absolute value of a positive ΔG is, the less stable a structure may be. Alterations in the most stable secondary structures of the sequences were visualized by comparing these structures with either common or rare alleles in parallel (Supplementary Figure S2).

In view of the fact that the length (800 bp) of sequence that the program can accept is shorter than the mature mRNA, the program does not account for the multiple mRNA-binding proteins that influence the conformation of the mature mRNA. Thus, this functional analysis is exploratory.

## RESULTS

Population stratification effects on associations were controlled after separating EAs and AAs in the analysis. The admixture degrees of our cleaned EA and AA samples were very low. They were 1.4% in EAs and 6.2% in AAs, respectively, and did not affect our association findings significantly (data not shown). As a result, in the present study, we only showed the *p*-values after these effects had been controlled.

The *p*-values ($<10^{-5}$) for the associations between the top-ranked SNPs and AD in the EA discovery sample are listed in Table 1, which includes 33 SNPs in 21 genes. All these SNPs in controls were in HWE in both EAs and AAs. Associations for five SNPs in *KIAA0040*, five SNPs in *SERINC2*, and one SNP in *HTR1A* in EAs had a genome-wide FDR < 0.05; eight of which were genome-wide significant ($p<5\times10^{-7}$) (Table 1). No associations for these SNPs (but other markers) were significant in AAs. The region surrounding *KIAA0040* in AAs overlapped extensively with that in EAs (Figure 1c). This region was enriched with risk or functional signals across EAs, AAs, six HapMap populations, and two primary tissues in Europeans.

This risk region spanned three LD blocks, including the first block from rs12094153 to rs1018829 in *TNN*, the second block from rs10489328 in *TNN* to rs1008459 in *KIAA0040* (38 kb), and the third block from rs2272785 to rs3753555 in *KIAA0040* (Figure 1l). The broader region (within 1 Mb; Figure 1b) outside this *TNN–KIAA0040* region yielded no association signal with $p<0.01$ in EAs. In all, 25 and 4 SNPs in this region were nominally associated with AD in EAs ($2.8\times10^{-7}\leqslant p\leqslant0.047$) and AAs ($5.1\times10^{-4}\leqslant p\leqslant0.040$), respectively. All association signals in this risk region with $p<10^{-5}$ in EAs were located in the second LD block in *KIAA0040* (Figure 1l). All risk SNPs were in HWE ($p>0.05$) in both cases and controls in both EAs and AAs (Table 2). Meta-analysis of EAs and AAs did not change the *p*-values for these risk SNPs significantly (data not shown). The allele frequencies of all SNPs in EA and AA controls were similar to those in CEU and YRI, respectively, in the HapMap database (see Supplementary Table S2).

eQTL analysis showed that all risk SNPs in this region had *cis*-acting regulatory effects on *KIAA0040* mRNA expression in at least one population; all of those with $p<10^{-5}$ in EAs had *cis*-acting effects in at least two populations; and five risk SNPs had significant effects in at least four populations (Table 2).

In EAs, throughout the Chromosome 1q, this region was the only one that had association signals at $p<10^{-5}$. Within 17.7 Mb around *TNN–KIAA0040*, this region was also the only one with association signals at $p<0.001$. Similarly, in AAs, within 10 Mb around *TNN–KIAA0040*, this region was the only one that had association signals at $p<0.001$. Furthermore, within 775 Kb around *TNN–KIAA0040*, this region was the only one with association signals at $p<0.01$. Additionally, within the 1 Mb range, the most significant functional SNPs in the HapMap CEU-Child (rs1057285; $p=0.0014$; Figure 1d), CEU-Parent (rs3737933; $p=0.004$; Figure 1e), YRI-Parent (rs1057302; $p=0.0002$; Figure 1f), and Europeans (brain tissue) (rs2269655; $p=0.0069$; Figure 1j) were all located in the second LD block, that is, *KIAA0040,* in this risk region (Figure 1l). The only exception was JPT, in which the most significant functional SNP in this risk region (rs4651322; $p=0.006$; Figure 1i) was located in the first LD block, that is, *TNN*; and this peak SNP was the second most significant SNP within the 1 Mb range.

The $-\log(p)$ values for all available SNPs ($n=40$) within *TNN–KIAA0040* region are plotted in Figure 1. The distributions of $-\log(p)$ values were highly consistent across EAs, AAs, CEU-Child, CEU-Parent, YRI-Child, YRI-Parent, and CHB ($0.369\leqslant r\leqslant0.824$; $2.8\times10^{-9}\leqslant p\leqslant0.032$; Table 3). The peak SNPs among each of these populations were in high LD. This was especially the case for the SNP showing the most significant expression differences in the CEU-Child population (rs1057285, $p=0.0014$), which was also the peak SNP associated with AD in AAs (rs1057285; $p=5.1\times10^{-4}$) (Figure 1c *vs* d). And the SNP showing the most significant expression differences in the YRI-Child population (rs1057239, $p=0.0108$) was the peak SNP associated with AD in EAs (rs1057239; $p=2.8\times10^{-7}$) (Figure 1b *vs* g). The peak SNPs in AAs (rs1057285), CEU-Child (rs1057285), CEU-Parent (rs3737933), and YRI-Parent (rs1057302) were closely located together (Figure 1l). The more closely the peak SNPs were located (Figure 1l), the more significant the correlations were between the $-\log(p)$ value distributions across the whole region (Table 2), which suggested that the peak SNP captured most of the information across that region. The more significant those correlations were, the more consistent (replicable) between two populations the risk regions were. Thus, the distance between peak SNPs reflected the strength of replicability of association or function signals between populations.

Transcriptome-wide *trans*-acting eQTL analysis showed that 10 SNPs in this region nominally regulated transcript expression of multiple genes across the transcriptome (Supplementary Table S3). Genome-wide *trans*-acting eQTL analysis showed that transcript expression of *KIAA0040* was marginally regulated by multiple genes across the genome (Supplementary Table S4). However, after Bonferroni correction, none of them remained significant.

Transcriptome-wide expression correlation analysis showed that the expression of *KIAA0040* was significantly correlated with the expression of many genes ($\alpha=1.7\times10^{-6}$; data not shown). These genes included some alcoholism-related genes (see Discussion), such as *SOD2* ($p=8.8\times10^{-11}$) and *ADH1C* ($p=1.2\times10^{-6}$) in brain, and *FAM44B* ($p<2\times10^{-16}$), *IPO11* ($p=1.9\times10^{-14}$), *ERAP1* ($p=1.2\times10^{-11}$), *GRIN2C* ($p=2.2\times10^{-11}$), *PECR* ($p=5.9\times10^{-11}$), *BBX* ($p=7.8\times10^{-11}$), *NRD1* ($p=9.7\times10^{-10}$),

**Table 3** Correlation of Distributions of –Log(*P*) Values for Associations and *Cis*-Acting Effects in *TNN–KIAA0040* Region Between Different Populations

| **Populations** | **Pearson correlation coefficients (*r*)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **EA** | **AA** | **CEU-child** | **CEU-parent** | **CHB** | **JPT** | **YRI-child** | **YRI-parent** | **Europeans**[a] |
| *P*-values | | | | | | | | | |
| EA | | 0.385 | 0.091 | 0.405 | −0.245 | −0.198 | 0.824 | 0.402 | −0.141 |
| AA | 0.025 | | 0.814 | 0.608 | 0.244 | 0.139 | 0.202 | 0.124 | 0.229 |
| CEU-child | 0.604 | $2.8 \times 10^{-9}$ | | 0.769 | 0.369 | 0.251 | 0.215 | 0.450 | 0.306 |
| CEU-parent | 0.014 | $8.3 \times 10^{-5}$ | $7.0 \times 10^{-8}$ | | 0.105 | 0.158 | 0.422 | 0.510 | 0.148 |
| CHB | 0.162 | 0.164 | 0.032 | 0.556 | | 0.215 | −0.190 | −0.036 | 0.090 |
| JPT | 0.270 | 0.440 | 0.158 | 0.380 | 0.229 | | −0.286 | 0.318 | −0.162 |
| YRI-child | $1.2 \times 10^{-8}$ | 0.268 | 0.244 | 0.018 | 0.306 | 0.125 | | 0.207 | 0.119 |
| YRI-parent | 0.028 | 0.507 | 0.013 | 0.004 | 0.848 | 0.093 | 0.264 | | −0.094 |
| Europeans | 0.552 | 0.331 | 0.190 | 0.532 | 0.707 | 0.509 | 0.627 | 0.710 | |

[a] Transcript expression in brain tissue.

*API5* ($p = 1.1 \times 10^{-9}$), *DRD2* ($p = 1.9 \times 10^{-9}$), *LEPR* ($p = 2.9 \times 10^{-9}$), *ADH5* ($p = 4.1 \times 10^{-9}$), *GRM5* ($p = 7.2 \times 10^{-9}$), *TH* ($p = 1.2 \times 10^{-8}$), *MTHFR* ($p = 1.6 \times 10^{-8}$), *CARS* ($p = 4.1 \times 10^{-8}$), *TTC12* ($p = 4.5 \times 10^{-8}$), *NXPH2* ($p = 2.2 \times 10^{-8}$), *CAST* ($p = 7.7 \times 10^{-8}$), *HNMT* ($p = 1.2 \times 10^{-7}$), *HTR1B* ($p = 1.7 \times 10^{-7}$), *OLFM3* ($p = 3.6 \times 10^{-7}$), *PPP1R1B* ($p = 7.1 \times 10^{-7}$), *OPRD1* ($p = 7.2 \times 10^{-7}$), *DRD3* ($p = 7.7 \times 10^{-7}$), *CRHR1* ($p = 1.3 \times 10^{-6}$), and *PPP2R2B* ($p = 1.5 \times 10^{-6}$) in PBMC.

Several risk SNPs were predicted to significantly alter the RNA secondary structure, including rs2157588 in *TNN*, and rs4650707, rs3737933, rs4650716, rs1894709, rs2861158, and rs16847872 in *KIAA0040* (Supplementary Figure S2 and Table S2). Eight *KIAA0040* markers were located in a transcription factor binding site (TFBS); two *KIAA0040* markers, that is, rs2861158 and rs2272784, were located in an exonic splicing silencer or enhancer (Supplementary Table S2).

## DISCUSSION

In the present study, after merging 480 COGA subjects into the SAGE sample, the results were highly similar to those in a previous study that used the SAGE sample alone (Bierut *et al*, 2010). The top-ranked risk SNPs ($p < 10^{-5}$) in EAs, AAs, and combined EAs and AAs in that previous study (Bierut *et al*, 2010) were confirmed by our analysis (see Supplementary Table S5). In the present study, we found genome-wide significant association signals ($p < 5 \times 10^{-7}$ together with FDR $< 0.05$) for AD for three genes (*KIAA0040, SERINC2* and *HTR1A*) in EAs. Two of these genes, that is, *KIAA0040* and *HTR1A*, were also among the top-ranked genes in EAs in the prior study (Bierut *et al*, 2010); in addition, *KIAA0040* as a risk gene for AD was confirmed by another GWAS meta-analysis for SAGE, COGA, and an Australian family sample (Wang *et al*, 2011). However, this was not previously replicated in AAs and not confirmed by functional studies.

In the present study, using a new analytic strategy and integrating evidence from the functional analysis, we were able to present additional important information that was not obtained previously. We found that, among the three significantly associated genes, only the region around *KIAA0040* overlapped extensively between EAs and AAs, which would be expected mostly for functional regions that harbor the same causal variant in both populations. This was consistent with the eQTL findings; that is, all risk SNPs in this region had expression effects. We thus concluded that *KIAA0040* might harbor a causal variant for AD.

Multiple pieces of evidence support our conclusion. First, *KIAA0040* contained two genome-wide significant markers and several other marginally significant markers in EAs. Second, *KIAA0040* was the only gene that had association signals with $p < 10^{-5}$ throughout Chromosome 1q in EAs. Similarly, in AAs, within 10 Mb around this region, *KIAA0040* was the only gene with association signals at $p < 0.001$. Furthermore, within the 1 Mb range, the most significant functional SNPs in four HapMap populations (Stranger *et al*, 2005) and one Duke sample (Heinzen *et al*, 2008) were all located in *KIAA0040*. It is thus likely that the putative causal variant for AD was located within *KIAA0040*. Third, eQTL analysis showed that all risk SNPs in this *TNN–KIAA0040* region had *cis*-acting regulatory effects, and such effects in *KIAA0040* appeared in two to five different populations, which is highly unlikely to have occurred by chance. These effects suggest that *KIAA0040 per se* might play a direct functional role in AD. Fourth, many *KIAA0040* SNPs had significant potential to alter the RNA secondary structures; many *KIAA0040* markers were located in a TFBS and two *KIAA0040* markers were located in an exonic splicing silencer or enhancer. Fifth, the overall –log(*P*) value distributions for gene–disease associations and for gene expression were correlated across at least seven populations, suggesting that the majority of the functions of *TNN–KIAA0040* might contribute to the risk for AD, and that the regulatory pathway through which these SNPs cause AD might be related to the TNN and KIAA0040 proteins *per se*.

In summary, (1) *TNN–KIAA0040* was enriched with many risk SNPs, (2) association signal distributions were consistent between EAs and AAs, (3) functional signal distributions were consistent across multiple populations, (4) association and functional signal distributions were consistent, and, especially, (5) many peak association and functional SNPs were concentrated in one LD block, suggesting that chance alone was unlikely to account for the findings. Taken together, these findings strongly support the hypothesis that *TNN–KIAA0040,* especially *KIAA0040,* might harbor a causal variant for AD.

*KIAA0040* encodes an HLA-DR11-restricted T-cell epitope. It is expressed in multiple tissues and organs including brain. It was worth noting that expression of *KIAA0040* was significantly associated with the expression of many genes that have previously been associated with AD (although some of these associations were reported by a candidate gene approach and were not yet well replicated). These genes are in the dopaminergic (*DRD2-TTC12, DRD3, TH*, and *PPP1R1B*), serotonergic (*HTR1B*), glutamatergic (*GRM5* and *GRIN2C*), histaminergic (*HNMT*), and opioidergic (*OPRD1)* systems, as well as in the ethanol metabolic pathway (*ADH1C* and *ADH5*) (Bierut *et al*, 2010; Edenberg *et al*, 2010; Connor *et al*, 2002; Yang *et al*, 2008; Dick and Foroud, 2003; Dahmen *et al*, 2005; Tabakoff *et al*, 2009; Sun *et al*, 2002; Oroszi *et al*, 2005; Zhang *et al*, 2008; Cichoz-Lach *et al*, 2007; Luo *et al*, 2006). These findings suggest that *KIAA0040* might also be implicated in AD via these neurotransmitter systems or metabolic pathways.

The causal variant is more likely to be located in *KIAA0040* than *TNN*, because (1) there were more risk SNPs in *KIAA0040* than *TNN*; (2) all risk SNPs with $p<10^{-5}$ in EAs were located in *KIAA0040*; (3) all functional markers that had significant *cis*-acting regulatory effects in at least two populations were located in *KIAA0040*; (4) all risk markers in AAs were located in *KIAA0040*; and (5) most peak association and functional SNPs were located in *KIAA0040*. However, most of these risk SNPs were common variants and were predicted to lack any phenotypic effect (by Polyphen-2; Adzhubei *et al*, 2010; Supplementary Table S2), so that the causal variant might not be any one of these risk markers *per se*. Future studies should aim to identify the causal variants by sequencing the entire *TNN–KIAA0040* region, especially *KIAA0040*.

*TNN* and *TNR* flank *KIAA0040*. They are closely linked, 8.9 and 129.7 kb distant from *KIAA0040*, respectively. *TNN*, which encodes tenascin-N, is involved in neurite outgrowth and cell migration in hippocampal explants. *TNR*, which encodes tenascin-R, is an extracellular matrix protein expressed primarily in the central nervous system and has been related to multiple brain diseases. Recent GWASs reported that (1) two SNPs in *KIAA0040* (rs1008459 and rs12136973) were significantly associated with amyotrophic lateral sclerosis (Schymick *et al*, 2007) and two other SNPs in *KIAA0040* (rs760486 and rs3766685) were marginally associated with Alzheimer's disease (Li *et al*, 2008); (2) several SNPs in *TNN* (rs1009418, rs12065394, rs6672099, rs6681984, and rs16847787) were marginally associated with narcolepsy, a neurological sleep disorder (Miyagawa *et al*, 2008); and (3) one inter-*KIAA0040–TNR* SNP (rs875326) was marginally associated with treatment response of schizophrenia to an antipsychotic medication (iloperidone) (Lavedan *et al*, 2009). These findings support a role for the *TNN–KIAA0040–TNR* compound locus in the risk for medical disorders, particularly those involving the central nervous system. In addition to the two aforementioned interpretations for our association findings in the present study (ie, *KIAA0040* might harbor a causal variant and directly contribute to the risk for AD, or it might be implicated in AD via neurotransmitter systems or metabolic pathways), these GWAS findings suggest to us an alternative interpretation that *KIAA0040* might regulate the risk for AD via flanking genes *TNN* or *TNR*.

The present study has limitations. First, the correction for multiple testing remains controversial. If corrected by Bonferroni correction ($\alpha=5\times10^{-8}$), which is conservative, our findings were only marginally significant. They warrant more validation independently in the future. Second, although $\lambda=1.07$ in EAs and $\lambda=1.04$ in AAs in QQ plots indicated that the inflation of $p$-values was not significant, we do not exclude the possibility that a small proportion of inflation might still exist, which might result from the heterogeneity of some unknown factors.

## ACKNOWLEDGEMENTS

This work was supported in part by National Institute on Drug Abuse (NIDA) Grants K01 DA029643, K24 DA017899, and K02 DA026990, National Institute on Alcohol Abuse and Alcoholism (NIAAA) Grants R01 AA016015, R21 AA020319, K24 AA013736, R01 AA11330, R01 AA017535, and P50 AA012870, and the National Alliance for Research on Schizophrenia and Depression (NARSAD) Award 17616 (to LZ). It also received support from the Department of Veterans Affairs through its support of the VA Alcohol Research Center, the VA National Center for PTSD, and the Depression REAP. We thank NIH GWAS Data Repository, the Contributing Investigator(s) (Drs Bierut and Edenberg) who contributed the phenotype and genotype data (SAGE and COGA) from his/her original study, and the primary funding organization that supported the contributing study. Funding support for SAGE and COGA was provided through the National Institutes of Health (NIH) Genes, Environment and Health Initiative (GEI) Grant U01 HG004422 and U01HG004438; the GENEVA Coordinating Center (U01 HG004446); the National Institute on Alcohol Abuse and Alcoholism (U10 AA008401); the National Institute on Drug Abuse (R01 DA013423); the National Cancer Institute (P01 CA089392); and the NIH contract 'High throughput genotyping for studying the genetic contributions to human disease' (HHSN268200782096C). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Genotyping was performed at the Johns Hopkins University Center for Inherited Disease Research or at deCODE.

## DISCLOSURE

Dr Kranzler has been a paid consultant for Alkermes, GlaxoSmithKline, and Gilead. He serves as a member of an Advisory Board for Lundbeck. He also reports associations with Eli Lilly, Janssen, Schering Plough, Lundbeck, Alkermes, GlaxoSmithKline, Abbott, and Johnson & Johnson, as these companies provide support to the ACNP Alcohol Clinical Trials Initiative (ACTIVE) and Dr Kranzler receives support from ACTIVE.

Dr Krystal has been a paid consultant for Aisling Capital, LLC, AstraZeneca Pharmaceuticals, Brintnall & Nicolini, Easton Associates, Gilead Sciences, GlaxoSmithKline, Janssen Pharmaceuticals, Lundbeck Research USA, Medivation, Merz Pharmaceuticals, MK Medical Communications, F Hoffmann-La Roche, SK Holdings, Sunovion Pharmaceuticals, Takeda Industries and Teva Pharmaceutical Industries. He serves as a member of Scientific Advisory Boards for Abbott Laboratories, Bristol-Myers Squibb, Eisai, Eli Lilly, Forest Laboratories, Lohocla Research Corporation, Mnemosyne Pharmaceuticals, Naurex, Pfizer Pharmaceuticals, and Shire Pharmaceuticals. He is the Editor for Biological Psychiatry, a member of Board of Directors of Coalition for Translational Research in Alcohol and Substance Use Disorders, and the President Elect for American College of Neuropsychopharmacology. He also gets support from Tetragenex Pharmaceuticals. The other authors declare no conflict of interest.

## REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al* (2010). A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248–249.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. 4th edn. American Psychiatric Press: Washington, DC.

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

Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E *et al* (2010). A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci USA* **107**: 5082–5087.

Cichoz-Lach H, Partycka J, Nesina I, Celinski K, Slomka M, Wojcierowski J (2007). Alcohol dehydrogenase and aldehyde dehydrogenase gene polymorphism in alcohol liver cirrhosis and alcohol chronic pancreatitis among Polish individuals. *Scand J Gastroenterol* **42**: 493–498.

Connor JP, Young RM, Lawford BR, Ritchie TL, Noble EP (2002). D(2) dopamine receptor (DRD2) polymorphism is associated with severity of alcohol dependence. *Eur Psychiatry* **17**: 17–23.

Dahmen N, Volp M, Singer P, Hiemke C, Szegedi A (2005). Tyrosine hydroxylase Val-81-Met polymorphism associated with early-onset alcoholism. *Psychiatr Genet* **15**: 13–16.

Dick DM, Foroud T (2003). Candidate genes for alcohol dependence: a review of genetic evidence from human studies. *Alcohol Clin Exp Res* **27**: 868–879.

Edenberg HJ, Koller DL, Xuei X, Wetherill L, McClintick JN, Almasy L *et al* (2010). Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcohol Clin Exp Res* **34**: 840–852.

Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M *et al* (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**: 944–947.

Grant BF, Dawson DA, Stinson FS, Chou SP, Dufour MC, Pickering RP (2004). The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991–1992 and 2001–2002. *Drug Alcohol Depend* **74**: 223–234.

Heath AC, Whitfield JB, Martin NG, Pergadia ML, Goate AM, Lind PA *et al* (2011). A quantitative-trait genome-wide association study of alcoholism risk in the community: findings and implications. *Biol Psychiatry* **70**: 513–518.

Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN *et al* (2008). Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol* **6**: e1.

Lavedan C, Licamele L, Volpi S, Hamilton J, Heaton C, Mack K *et al* (2009). Association of the NPAS3 gene and five other loci with response to the antipsychotic iloperidone identified in a whole genome association study. *Mol Psychiatry* **14**: 804–819.

Li H, Wetten S, Li L, St Jean PL, Upmanyu R, Surh L *et al* (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol* **65**: 45–53.

Luo X, Kranzler HR, Zuo L, Wang S, Schork NJ, Gelernter J (2006). Diplotype trend regression analysis of the ADH gene cluster and the ALDH2 gene: multiple significant associations with alcohol dependence. *Am J Hum Genet* **78**: 973–987.

Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A *et al* (2008). Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. *Nat Genet* **40**: 1324–1328.

Oroszi G, Enoch MA, Chun J, Virkkunen M, Goldman D (2005). Thr105Ile, a functional polymorphism of histamine N-methyltransferase, is associated with alcoholism in two independent populations. *Alcohol Clin Exp Res* **29**: 303–309.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–575.

Schumann G, Coin LJ, Lourdusamy A, Charoen P, Berger KH, Stacey D *et al* (2011). Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proc Natl Acad Sci USA* **108**: 7119–7124.

Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, Gibbs JR *et al* (2007). Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* **6**: 322–328.

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R *et al* (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**: e78.

Sun HF, Chang YT, Fann CS, Chang CJ, Chen YH, Hsu YP *et al* (2002). Association study of novel human serotonin 5-HT(1B) polymorphisms with alcohol dependence in Taiwanese Han. *Biol Psychiatry* **51**: 896–901.

Tabakoff B, Saba L, Printz M, Flodman P, Hodgkinson C, Goldman D *et al* (2009). Genetical genomic determinants of alcohol consumption in rats and humans. *BMC Biol* **7**: 70.

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

Treutlein J, Cichon S, Ridinger M, Wodarz N, Soyka M, Zill P *et al* (2009). Genome-wide association study of alcohol dependence. *Arch Gen Psychiatry* **66**: 773–784.

Wang KS, Liu X, Zhang Q, Pan Y, Aragam N, Zeng M (2011). A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence. *J Psychiatr Res*; doi:10.1016/j.jpsychires.2011.06.005.

Yang BZ, Kranzler HR, Zhao H, Gruen JR, Luo X, Gelernter J (2008). Haplotypic variants in DRD2, ANKK1, TTC12, and NCAM1 are associated with comorbid alcohol and drug dependence. *Alcohol Clin Exp Res* **32**: 2117–2127.

Zhang H, Kranzler HR, Yang BZ, Luo X, Gelernter J (2008). The OPRD1 and OPRK1 loci in alcohol or drug dependence: OPRD1 variation modulates substance dependence risk. *Mol Psychiatry* **13**: 531–543.

Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

Supplementary Information accompanies the paper on the Neuropsychopharmacology website (http://www.nature.com/npp)