Exome-Wide Association Study of Replicable Nonsynonymous Variants Conferring Risk for Alcohol Dependence

LINGJUN ZUO, M.D.,[a,*] LAURA SABA, PH.D.,[b] KESHENG WANG, PH.D.,[c] XIANGYANG ZHANG, M.D., PH.D.,[d] JOHN H. KRYSTAL, M.D.,[a] BORIS TABAKOFF, PH.D.,[b] AND XINGGUANG LUO, M.D., PH.D.[a,*]

[a]*Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut*

[b]*Department of Pharmacology, University of Colorado Denver School of Medicine, Aurora, Colorado*

[c]*Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, Tennessee*

[d]*Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, Texas*

ABSTRACT. Objective: In the present study, we scanned the whole exome in three independent samples to search for replicable risk nonsynonymous (ns) variants (ns single-nucleotide polymorphisms [nsSNPs]) for alcohol dependence. **Method:** A total of 10,554 subjects in three independent samples were analyzed for association with alcohol dependence, including one European American sample (1,409 cases with alcohol dependence and 1,518 controls), one African American sample (681 cases and 508 controls), and one European Australian sample (a total of 6,438 family subjects with 1,645 alcohol-dependent probands). RNA expression of the risk genes in human, mouse, and rat brains was also explored. **Results:** We identified a total of 70 nsSNPs at 65 genes with nominally replicable associations; 22 nsSNPs at 21 genes among them survived corrections for multiple testing in meta-analysis ($\alpha = .0007$). By incorporating the information from bioinformatics and RNA expression analyses, we identified at least two of the most promising risk genes for alcohol dependence: *APOER2* and *UBAP2.* **Conclusions:** The gene coding for apolipoprotein E receptor 2 (*APOER2*) and the gene coding for ubiquitin-associated protein-2 (*UBAP2*) are among the most appropriate for follow-up in human and nonhuman species as contributors to risk for alcohol dependence. (*J. Stud Alcohol Drugs, 74,* 622–625, 2013)

ALTHOUGH PERCEPTIONS MAY CHANGE, across the genome most variants (>98.5%) are currently thought of as "silent" mutations that include variants in the intergenic or intronic regions (i.e., noncoding regions) and synonymous variants in the exonic regions (i.e., coding regions). Many of these silent mutations have been associated with susceptibility to human diseases. However, one important interpretation of these associations is that these silent mutations might be in linkage disequilibrium with nonsynonymous (ns) variants (ns single-nucleotide polymorphisms [nsSNPs]) in the coding regions that are more likely to be functional and thus disease causal. It is estimated that the protein coding regions of the human genome constitute about 85% of the disease-causing mutations (Choi et al., 2009). In the present study, we scanned the whole exome to search for risk nsSNPs for alcohol dependence whose associations were replicable in multiple populations and whose functions were validated by multiple approaches, including bioinformatics analysis, *cis*-acting expression quantitative trait locus (*cis*-eQTL) analysis in human brains, and RNA expression analysis in mouse and rat brains. Replication and validation reduced the chance of false positives. Instead of correction by the number of all nsSNPs, the replicable associations needed correction by only the number of replicable risk markers.

Method

A total of 10,554 subjects in three independent samples were analyzed for association with alcohol dependence (according to the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*; American Psychiatric Association, 1994), including one European American sample (1,409 cases with alcohol dependence and 1,518 controls; from the Study of Addiction Genetics and Environment [SAGE] and Collaborative Study on the Genetics of Alcoholism [COGA] data sets in the database of Genotypes and Phenotypes), one African American sample (681

Received: November 30, 2012. Revision: February 6, 2013.

This work was supported by National Institute on Drug Abuse Grant K01 DA029643, National Institute on Alcohol Abuse and Alcoholism Grant R21 AA020319, and National Alliance for Research on Schizophrenia and Depression Award 17616 (to Lingjun Zuo). The authors thank the National Institutes of Health Genome-Wide Association Studies Data Repository (database of Genotypes and Phenotypes); the contributing investigator(s) (Drs. Bierut, Edenberg, and Heath) who contributed the phenotype and genotype data (SAGE: phs000092.v1.p1, COGA: phs000125.v1.p1 and OZ-ALC: phs000181.v1.p1) from his/her original study; and the funding support by U01 HG004422, U01HG004438, U01 HG004446, U10 AA008401, R01 DA013423, and HHSN268200782096C.

*Correspondence may be sent to Lingjun Zuo and Xingguang Luo at the Yale University School of Medicine, West Haven, CT 06516, or via email at: Lingjun.Zuo@yale.edu and Xingguang.Luo@yale.edu.

cases and 508 controls; from the SAGE and COGA data sets), and one European Australian sample (a total of 6,438 family subjects with 1,645 alcohol-dependent probands; from the OZ-ALC [the Australian twin family study of alcohol use disorder] data set). Detailed demographic information for these samples has been published previously (Bierut et al., 2010; Edenberg et al., 2010; Heath et al., 2011; Zuo et al., 2011, 2012, 2013). The European American and African American samples were genotyped on the Illumina Human 1M BeadChip (Illumina, Inc., San Diego, CA), and the Australian sample was genotyped on the Illumina CNV370v1 BeadChip (Illumina, Inc., San Diego, CA).

Before association analysis, we stringently cleaned the phenotype and genotype data of these three samples. Excluded were subjects with poor genotypic data; subjects with allele discordance, sample relatedness, gender anomalies, chromosome anomalies (such as aneuploidy and mosaic cell populations), missing race, and non-European and non-African ancestries; subjects with a mismatch between self-identified and genetically inferred ethnicity; and subjects with a missing genotype call rate greater than 2% across all SNPs. Furthermore, SNPs with an allele frequency difference in controls greater than 2% between SAGE and COGA, SNPs with a missing rate difference greater than 2% between SAGE and COGA, and SNPs with allele discordance were excluded. We then filtered out the SNPs on all chromosomes with an overall missing genotype call rate greater than 2% and the SNPs with minor allele frequencies less than 0.01 in all populations examined. SNPs that deviated from the Hardy–Weinberg equilibrium ($p < 10^{-4}$) within controls were also excluded. A total of 14,657 cleaned variants extracted from 19,504 nsSNPs were tested for associations with alcohol dependence using the logistic regression analysis implemented in PLINK (for case-control samples) or using the family-based association test implemented in FBAT (for family-based samples) (Zuo et al., 2011).

Furthermore, we performed *cis*-eQTL analysis of detected risk variants. Expression data were evaluated in 93 European, autopsy-collected, frontal cortical brain tissue samples with no defined neuropsychiatric condition (Heinzen et al., 2008). Differences in the distribution of mRNA expression levels between SNP genotypes were compared using a linear regression model that corrected for age and sex. Finally, RNA expression of the risk genes in mouse and rat brains was also explored using the Affymetrix Mouse (Rat) Exon 1.0 ST array and RNA-Seq technology (Trapnell et al., 2010).

Results

We found 807, 671, and 279 nsSNPs that were nominally associated with alcohol dependence in European Americans, African Americans, and European Australians, respectively ($p < .05$; data not shown). The top-ranked SNPs were rs11120301 (at *SMYD2;* $p = 1.1 \times 10^{-5}$), rs3820198 (at *APOER2;* $p = 2.7 \times 10^{-5}$), and rs961360 (at *R3HDM1;* $p = 2.5 \times 10^{-5}$) in these three populations, respectively (Table 1), which were suggestively associated with alcohol dependence after correction for exome-wide multiple testing ($\alpha = 3.4 \times 10^{-6}$). Additionally, 36, 14, and 24 associations were nominally replicable between European Americans and African Americans, between European Americans and European Australians, and between African Americans and European Australians, respectively (all $p < .05$).

Two associations were nominally replicable across three populations (rs1052439 at *FAM79B* and rs7927370 at *OR4A15*). Among a total of 70 nsSNPs at 65 genes with nominally replicable associations, 22 survived corrections for multiple testing in meta-analysis and were thus taken as significant and replicable ones ($\alpha = .0007$; Table 1). Among these 22 replicable nsSNPs, 18 nsSNPs with similar minor allele frequencies between different populations had the same directions of gene effects across the three samples. Four nsSNPs (i.e., rs3820198, rs2125579, rs12075, and rs7627615) with a significant difference in minor allele frequencies between different populations had the opposite directions of gene effects between European Americans/Australians and African Americans. Ten nsSNPs were located at the exonic splicing enhancer or exonic splicing silencer. Four nsSNPs (at *RSNL2, UBAP2, ANKRD30A,* and *RPAP1*) were predicted to affect protein function or structure (possibly damaging). Four nsSNPs (at *APOER2, DARC, TRPM6,* and *HTR3E*) have been reported to be directly associated with other medical diseases or traits, including one with schizophrenia (i.e., rs7627615 at *HTR3E*) (Lennertz et al., 2010) and one with Parkinson's disease (i.e., rs3820198 at *APOER2*) (Chen et al., 2012). Thirteen nsSNPs affected mRNA expression of local genes (*cis*-eQTL) in the human brain ($3.7 \times 10^{-4} < p < .05$).

Additionally, among these replicable genes, expression of 14 and 14 genes could be distinguished above background (all $p < .0001$) in whole-brain samples of all 353 young adult male mice (62 inbred strains from the ILSXISS recombinant inbred panels including the parental strains) and in 108 young adult male rats (27 inbred strains from the HXB/BXH recombinant inbred panels and its related inbred strains), respectively. Expression of 6 and 3 genes had an RPKM (reads per kilobase per million) of 5.0 or greater in mouse and rat brains, respectively.

Discussion

The main goal of an association study is to pinpoint the disease causal variants. Therefore, it is a promising strategy to study nsSNPs that are more likely to be functional,

TABLE 1. Top-ranked or replicable nsSNPs associated with alcohol dependence

SNP	CHR	Gene	Amino acid	*p* values for associations (risk allele frequency: affected/unaffected)			Meta	*cis*-eQTL	Splicing	Damaging effects	Diseases /traits[#]	Mouse* (RPKM)▲	Rat* (FPKM)◆
				EA	AA	EAu							
rs3820198●	1	*APOER2*	D45E	0.015 (T: .681/.650)	2.7×10^{-5} (G: .869/.811)	–	5.9×10^{-5}	0.013	Y	benign	1	7.17	N.A.
rs3820678	1	*OR10R2*	A190T	0.008 (A: .223/.195)	–	0.027 (A: .260/.233)	7.0×10^{-4}	0.273	–	benign		N.A.	N.A.
rs12075	1	*DARC*	G41D	0.007 (A: .599/.563)	0.029 (G: .145/.131)	–	7.0×10^{-4}	N.A.	–	benign	2	3.85	N.A.
rs12118628	1	*OR10J1*	M111I	0.002 (G: .894/.867)	–	0.029 (G: .876/.857)	2.0×10^{-4}	0.038	–	benign		0.00	N.A.
rs11120301●	1	*SMYD2*	I429M	1.1×10^{-5} (G: .010/.003)	0.280 (G: .005/.002)	N.A.	9.5×10^{-6}	N.A.	Y	benign		15.95	18.59
rs3100246	2	*RSNL2*	R485L	0.011 (G: .854/.831)	0.003 (G: .981/.959)	–	3.0×10^{-4}	0.049	Y	damaging		4.50	0.00
rs961360●	2	*R3HDM1*	M269V	0.441 (T: .838/.822)	0.311 (T: .802/.784)	2.5×10^{-5} (T: .904/.872)	1.0×10^{-4}	N.A.	Y	damaging		16.80	24.46
rs7627615	3	*HTR3E*	A85T	–	0.027 (A: .882/.859)	0.003 (G: .413/.377)	3.0×10^{-4}	0.047	Y	benign	3	N.A.	N.A.
rs1052439	3	*FAM79B*	L3P	0.045 (T: .673/.647)	0.038 (T: .659/.617)	0.006 (T: .700/.669)	1.0×10^{-4}	0.087	Y	unknown		N.A.	N.A.
rs16902872	5	*FLJ25422*	A270V	0.003 (A: .003/.001)	0.016 (A: .290/.254)	N.A.	2.0×10^{-4}	N.A.	Y	benign		4.41	N.A.
rs846664	7	*TAS2R16*	N171K	0.005 (G: .004/.001)	0.041 (G: .286/.252)	N.A.	6.0×10^{-4}	N.A.	–	benign		0.00	0.00
rs307658	9	*UBAP2*	N605S	–	0.019 (A: .495/.449)	0.007 (A: .642/.606)	7.0×10^{-4}	6.1×10^{-4}	–	benign		N.A.	N.A.
rs1785506	9	*UBAP2*	R13Q	–	0.019 (A: .494/.448)	0.006 (A: .641/.605)	6.0×10^{-4}	6.1×10^{-4}	YY	damaging		93.98	6.79
rs3750425	9	*TRPM6*	V1392I	–	0.046 (T: .281/.246)	2.6×10^{-5} (T: .102/.073)	3.3×10^{-6}	0.002	–	benign	4	0.12	N.A.
rs1200875	10	*ANKRD30A*	R928C	0.022 (G: .781/.754)	–	0.009 (G: .782/.754)	5.0×10^{-4}	0.048	–	damaging		N.A.	N.A.
rs2792751	10	*GPAM*	I42V	0.021 (A: .304/.279)	–	0.012 (A: .287/.261)	7.0×10^{-4}	3.7×10^{-4}	Y	benign		7.53	2.99
rs1541314	11	*TOLLIP*	G1807S	0.015 (A: .084/.069)	–	0.006 (A: .072/.055)	2.0×10^{-4}	0.23	–	benign		13.14	34.80
rs2273549	11	*TCP11L1*	K177R	0.037 (A: .835/.815)	–	0.005 (A: .834/.806)	5.0×10^{-4}	0.025	–	benign		7.37	N.A.
rs7927370	11	*OR4A15*	A286V	0.044 (C: .952/.939)	0.044 (C: .993/.984)	0.035 (C: .948/.933)	7.0×10^{-4}	0.413	–	benign		0.00	0.00
rs11630901	15	*RPAP1*	R581G	0.005 (T: .830/.804)	0.023 (T: .972/.957)	–	4.0×10^{-4}	0.123	Y	damaging		1.77	3.18
rs1139897	16	*RHOT2*	R244Q	–	0.007 (G: .945/.921)	0.009 (G: .783/.753)	6.0×10^{-4}	0.016	Y	benign		N.A.	9.91
rs1800309	17	*GAA*	E688K	0.008 (G: .971/.957)	0.002 (G: .986/.967)	N.A.	2.0×10^{-4}	N.A.	Y	benign		33.61	N.A.
rs4806163	19	*ZD52F10*	V90A	–	0.012 (A: .054/.031)	0.008 (A: .232/.204)	6.0×10^{-4}	0.016	–	benign		0.63	0.50
rs2125579	19	*ZNF235*	H295P	0.008 (G: .559/.529)	0.060 (T: .808/.780)	0.044 (G: .549/.522)	2.0×10^{-4}	0.031	–	benign		N.A.	N.A.

●These three single-nucleotide polymorphisms (SNPs) are top-ranked SNPs in three samples, respectively; two of them are not replicable. Risk allele = the allele having higher frequency in cases/transmitted than controls/untransmitted; EA = European Americans; AA = African Americans; EAu = European Australians; *cis*-eQTL = minimal *p* values for exon-level *cis*-acting expression regulation in human brain; Y = located at exon-splicing enhancer (ESE) or exon-splicing silencer (ESS); YY = located at ESE or ESS that can abolish protein domain. [#]These diseases/traits are associated with nsSNPs: (1) Parkinson's disease, plasma cholesterol levels as well as size and composition of low-density lipoprotein particles; (2) Inflammation, white blood cell count, circulating concentrations of monocyte chemoattractant protein-1; (3) Schizophrenia; (4) Type 2 diabetes.*Transcript-level RNA expression in mouse and rat brains; raw and processed data are available at http://phenogen.ucdenver.edu; expression of all genes with RPKM > 0 listed in this table could be distinguished above background ($p < .0001$) in whole brain samples of all mice and rats examined using the Affymetrix Mouse (Rat) Exon 1.0 ST array. ▲RPKM = reads per kilobase of transcript per million mapped reads; these RPKM values were calculated from the RNA-Seq data generated from ribosomal depleted total RNA from six mouse brains using the Helicos Helisphere single molecule sequencing system. ◆FPKM = fragments per kilobase of transcript per million mapped reads; these FPKM values were calculated from the RNA-Seq data generated from polyA+ selected RNA from three rat brains using the Illumina HiSeq2000. "–", $p > .05$; N.A. = not available.

have larger effect sizes, and have a higher penetrance than silent mutations. Theoretically, all of the top-ranked or replicable genes noted above might contribute to the risk of alcohol dependence because it is a multigenic disorder. Among these nsSNPs, variants that are more likely to be functional would be the higher priority. Usually, the causal variants have stronger associations with diseases than non-causal markers; thus, those top-ranked variants (at *SMYD2, APOER2,* and *R3HDM1*) and those risk variants surviving correction for multiple testing are more likely to be causal. Further, variants that were replicable across multiple populations, especially across those genetically distinct populations such as European Americans and African Americans, are more likely to be functional. Among these replicable nsSNPs, two nsSNPs (at *FAM79B* and *OR4A15*) were replicable across three populations. Variants with significant *cis*-acting regulatory effects on gene expression in the human brain are also more likely to be functional. Among the 13 replicable nsSNPs with nominally significant *cis*-regulatory effects, four nsSNPs at three genes (*UBAP2, TRPM6,* and *GPAM*) remained significant even after correction for the numbers of exons within each risk gene and the number of nsSNPs examined.

Variants that are located at the exonic splicing enhancer or exonic splicing silencer may disrupt splicing activity and cause alternative splicing, especially the one at *UBAP2* that could abolish a protein domain. The four nsSNPs (at *RPAP1, RSNL2, ANKRD30A,* and *UBAP2*) that were predicted to affect protein function or structure (possibly damaging) are also likely to be functional. Four nsSNPs reported to be directly associated with other medical diseases or traits, especially the two with brain disorders (at *HTR3E* and *APOER2*), are of substantial interest. Finally, the human candidate genes that are expressed in the brains of other species as shown in Table 1 can be the starting point for much more detailed testing of hypotheses generated by our studies with humans. Integrating all of the above rationale, we believe that gene coding for apolipoprotein E receptor 2, *APOER2,* and the gene coding for ubiquitin-associated protein-2, *UBAP2,* are among the most appropriate for follow-up in human and nonhuman species as contributors to risk for alcohol dependence.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bierut, L. J., Agrawal, A., Bucholz, K. K., Doheny, K. F., Laurie, C., Pugh, E., . . . Rice, J. P., as part of the Gene, Environment Association Studies (GENEVA) Consortium. (2010). A genome-wide association study of alcohol dependence. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 5082–5087.

Chen, K., Chen, Y. P., Song, W., Huang, R., Zhao, B., Cao, B., . . . Shang, H.-F. (2012). Association analysis of *LRP8* SNP rs3820198 and rs5174 with Parkinson's disease in Han Chinese population. *Neurological Research, 34,* 725–729.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., . . . Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 19096–19101.

Edenberg, H. J., Koller, D. L., Xuei, X., Wetherill, L., McClintick, J. N., Almasy, L., . . . Foroud, T. (2010). Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcoholism: Clinical and Experimental Research, 34,* 840–852.

Heath, A. C., Whitfield, J. B., Martin, N. G., Pergadia, M. L., Goate, A. M., Lind, P. A., . . . Montgomery, G. W. (2011). A quantitative-trait genome-wide association study of alcoholism risk in the community: Findings and implications. *Biological Psychiatry, 70,* 513–518.

Heinzen, E. L., Ge, D., Cronin, K. D., Maia, J. M., Shianna, K. V., Gabriel, W. N., . . . Goldstein, D. B. (2008). Tissue-specific genetic control of splicing: Implications for the study of complex traits. *PLoS Biology, 6*(12), e1000001.

Lennertz, L., Wagner, M., Frommann, I., Schulze-Rauschenbach, S., Schuhmacher, A., Kühn, K.-U., . . . Mössner, R. (2010). A coding variant of the novel serotonin receptor subunit 5-HT3E influences sustained attention in schizophrenia patients. *European Neuropsychopharmacology, 20,* 414–420.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28,* 511–515.

Zuo, L., Gelernter, J., Zhang, C. K., Zhao, H., Lu, L., Kranzler, H. R., . . . Luo, X. (2012). Genome-wide association study of alcohol dependence implicates *KIAA0040* on chromosome 1q. *Neuropsychopharmacology, 37,* 557–566.

Zuo, L., Zhang, C. K., Wang, F., Li, C.-S., Zhao, H., Lu, L., . . . Luo, X. (2011). A novel, functional and replicable risk gene region for alcohol dependence identified by genome-wide association study. *PLoS ONE, 6*(11), e26726.

Zuo, L., Zhang, H., Malison, R. T., Li, C.-S. R., Zhang, X.-Y., Wang, F., . . . Luo, X. (2013). Rare *ADH* variant constellations are specific for alcohol dependence. *Alcohol and Alcoholism, 48,* 9–14.