Published in final edited form as:
Int J Neuropsychopharmacol. 2013 March ; 16(2): . doi:10.1017/S1461145712000375.

Genome-Wide Gene Set Analysis for Identification of Pathways Associated with Alcohol Dependence

Joanna M. Biernacka[§,1,2], **Jennifer Geske**[1], **Gregory D. Jenkins**[1], **Colin Colby**[1], **David N. Rider**[1], **Victor M. Karpyak**[2], **Doo-Sup Choi**[2,3], and **Brooke L. Fridley**[1]

[1]Divisions of Biomedical Statistics & Informatics, Department of Health Sciences Research

[2]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN 55905, USA

[3]Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN 55905, USA

Abstract

It is believed that multiple genetic variants with small individual effects contribute to the risk of alcohol dependence. Such polygenic effects are difficult to detect in genome-wide association studies that test for association of the phenotype with each single nucleotide polymorphism (SNP) individually. To overcome this challenge, gene set analysis (GSA) methods that jointly test for the effects of pre-defined groups of genes have been proposed. Rather than testing for association between the phenotype and individual SNPs, these analyses evaluate the global evidence of association with a set of related genes enabling the identification of cellular or molecular pathways or biological processes that play a role in development of the disease. It is hoped that by aggregating the evidence of association for all available SNPs in a group of related genes, these approaches will have enhanced power to detect genetic associations with complex traits. We performed GSA using data from a genome-wide study of 1165 alcohol dependent cases and 1379 controls from the Study of Addiction: Genetics and Environment (SAGE), for all 200 pathways listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Results demonstrated a potential role of the “Synthesis and Degradation of Ketone Bodies” pathway. Our results also support the potential involvement of the “Neuroactive Ligand Receptor Interaction” pathway, which has previously been implicated in addictive disorders. These findings demonstrate the utility of GSA in the study of complex disease, and suggest specific directions for further research into the genetic architecture of alcohol dependence.

Keywords

pathway; gene set analysis; alcohol dependence; alcoholism; ketone bodies; neuroactive ligand receptors

Introduction

Alcohol dependence and related alcohol use disorders are known to be under considerable genetic influence (Agrawal and Lynskey, 2008; Goldman et al., 2005); yet few specific genetic risk factors have been discovered and confirmed (Ball, 2008). In recent years, genome-wide association studies (GWAS) have emerged as a powerful approach to identify

[§]Corresponding Author: Joanna M. Biernacka, Ph.D., Department of Health Science Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Tel: 507-538-5274; Fax: 507-284-9542; biernacka.joanna@mayo.edu.

Statement of Interest: None.

disease susceptibility genes (Kronenberg, 2008; Lettre and Rioux, 2008). In these studies, the association of a phenotype with hundreds of thousands of single nucleotide polymorphisms (SNPs) distributed throughout the genome is evaluated. This approach has been applied to the study of many complex traits, including alcohol dependence (Bierut et al., 2010; Treutlein et al., 2009). Despite the potential of GWAS to identify novel genetic contributors to complex diseases, these studies tend to be greatly underpowered for detection of disease-related SNPs with small or moderate effect sizes. Thus, GWAS have had limited success in identifying genes that influence risk of alcohol dependence.

Phenotypic characteristics are believed to be controlled by networks of interacting biochemical and physiological pathways influenced by the products of many genes. While single genetic variants may only have a small influence on complex traits, the combined effects of genes within a biochemical pathway have a greater potential to impact complex phenotypes. Recent studies suggest that assessing the effects of genetic variants in the context of pathways is critical to understanding their phenotypic significance (Srinivasan et al., 2009; Wang et al., 2007). However, most GWAS rely on analyses of individual SNPs, which ignore prior knowledge about gene function and role of genes within molecular pathways. To overcome this limitation, gene set analysis (GSA) methods for genome-wide SNP data have recently been introduced (Fridley and Biernacka, 2011; Holmans, 2010; Wang et al., 2010). GSA incorporates prior biological knowledge into statistical analysis by evaluating the overall evidence of association of a phenotype with all genotyped SNPs in a pre-specified set of genes defined, for example, based on their role in a particular molecular pathway. Such methods may enable the detection of subtle effects of multiple genes in the same gene set that may be missed by assessing each SNP or gene individually. Moreover, the incorporation of biological knowledge in GSA may aid researchers in the interpretation of results and help focus further research efforts.

GSA, which has also been referred to as pathway analysis or functional gene group analysis, has recently been applied to a number of neuropsychiatric phenotypes including cognitive ability (Ruano et al., 2010) and bipolar disorder (Holmans et al., 2009; O'Dushlaine et al., 2010). This paper describes gene-set analyses of genome-wide data from the Study of Addiction: Genetics and Environment (SAGE) (Bierut et al., 2010), for all pathways listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.jp/kegg/). Recently developed methods were applied to perform these pathway-based analyses (Biernacka et al., 2011). The goal of these analyses was to identify sets of genes that contribute to individual differences in susceptibility to alcohol dependence and thereby identify new candidates for ongoing studies.

Methods

Data

Genome-wide SNP and phenotypic data were downloaded from the database on Genotypes and Phenotypes (dbGaP; study accession phs000092.v1.p1). These data were collected by SAGE, which is part of the NIH-funded Gene Environment Association Studies initiative (GENEVA). Description of the study design, subjects, and results of genome-wide association analyses have been published (Bierut et al., 2010). Briefly, 1944 alcohol dependent cases and 1965 controls were genotyped with the Illumina Human1Mv1_C BeadChips. All subjects were phenotypically assessed using the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA). This study of the association of gene sets with alcoholism focused on the European American subset of subjects. Following the exclusion of non-European American subjects, quality control filters for the European American subject set distributed with the data were applied. The quality control included a data-cleaning step recommended by the SAGE investigators that consisted of setting the genotypes in specific regions for several individuals to missing. These regions were identified as anomalous genotype intensity patterns which may indicate aneuploidy or problems with genotyping. The quality control filters also excluded samples with missing call rate ≥ 2%, and SNPs with missing call rate ≥ 2%, minor allele frequency (MAF) < 1%, or Hardy-Weinberg Equilibrium P-value < 10^{-4}. After applying the quality control filters established by the SAGE investigators, a further 209 SNPs were removed as they had minor allele frequencies below 1% in the final subset of subjects, and 1257 SNPs were removed because of evidence of departure from Hardy Weinberg Equilibrium (p<0.001). The final data set consisted of 839,409 SNPs in 2544 subjects, including 1165 alcohol dependent cases and 1379 controls.

Prior to performing the association analyses, we analyzed the data using EIGENSTRAT (Price et al., 2006) to determine principal components (PCs) that capture any remaining population stratification among the European American subjects. The Tracy-Widom test was used to test for eigenvalues > 0, and determined that three PCs were needed to adjust for population stratification in this subset of subjects; these PCs were used as covariates in the gene set association analyses.

Definition of Gene-Sets

Lists of genes corresponding to all pathways in the KEGG database as of February 19, 2010 (Version 53.0) were generated. SNPs were then assigned to genes by including SNPs that map within 5kb upstream or downstream of the first and last exon of the gene. This algorithm led to assigning 127,047 SNPs to 200 gene sets representing pathways annotated in the KEGG database. These pathways and their characteristics (number of genes and number of SNPs) are listed in Supplemental Table S1. We note that five of the analyzed gene sets were previously reported to play a role in alcohol and drug addiction (Li et al., 2008) [KARG; http://karg.cbi.pku.edu.cn/]. These gene sets represent the “neuroactive ligand-receptor interaction”, “long-term potentiation”, “GnRH signaling”, ”MAPK signaling”, and “Gap junctions” pathways.

Statistical Analysis

We used two approaches for GSA, a one-step analysis method and a two-step method. In the one-step approach all SNPs in a gene set are used in the analysis without consideration of gene-level effects. In the two-step approach all genotyped SNPs in each gene are first used to evaluate association with the gene, followed by aggregation of the gene-level tests to assess evidence for association of the phenotype with the gene set (Biernacka et al., 2011; Fridley and Biernacka, 2011).

It has been demonstrated that Fisher's method (Fisher, 1932) is a powerful approach for combining SNP-specific p-values in the context of gene-set analyses (Chai et al., 2009). Fisher's method is based on transforming a set of p-values, and then adding the transformed p-values to obtain a test statistic. Rather than applying the -2log(p) transformation of Fisher's method, other transformations of p-values can be applied. One way of combining p-values is the Gamma method, which is also based on summing transformed p-values, but uses an inverse Gamma transformation (Zaykin et al., 2007). For a particular shape parameter ω, the test statistic is defined as $\sum_{i=1}^{N} G_{\omega,1}^{-1}(1-p_i)$, where G^{-1} is the inverse of a Gamma(ω, 1) cumulative distribution function (Zaykin et al., 2007). It can be shown that Fisher's method is a special case of Gamma method, with a shape parameter of $\omega = 1$. We have demonstrated that for SNP-based GSA, the Gamma method with a small shape parameter, in the range of 0.1-0.2, can be more powerful than Fisher's method (Biernacka et al., 2011). Therefore, for our one-step GSA, we used the Gamma method (GM) with shape parameter 0.1 ($GM_{0.1}$). However, the optimal shape parameter in a GM GSA may depend on factors such as the number of SNPs and linkage disequilibrium (LD) structure within the gene set, as well as the true underlying disease-causing model. Moreover, we have previously shown that combination of p-values from univariate analyses using FM (Fisher, 1932), with permutation-based assessment of significance of association with the gene-set, is a powerful approach for GSA of expression data (Fridley et al., 2010). Therefore, as a sensitivity analysis, we also analyzed the data using the GM with a shape parameter of 1 (GM_1), which is equivalent to the commonly used Fisher's method (FM). Results of these secondary analyses are shown along with the results of the main analysis in Supplemental Table S1.

For the one-step approach, association tests were first performed for each SNP individually using logistic regression, with alcohol dependence as the response variable. For these analyses, SNP genotypes were coded as the dosage of minor alleles, which assumes log-additive allele effects. For each gene set, the gene set statistic was then calculated using the GM with shape parameter 0.1 ($GM_{0.1}$), and as a sensitivity analysis with a shape parameter of 1 (GM_1). Rather than using the asymptotic distribution of the gene set test statistic, which assumes independence of SNP-specific p-values, permutations were used to assess significance of the gene-set statistic. Because the permutation method that was applied maintains the LD structure in the data, this method of calculating the gene-set p-value correctly accounts for the observed LD between SNPs.

In the two-step approach, all genotyped SNPs in a gene were first used to evaluate the phenotypic association with each gene, and the gene-level tests were then aggregated to assess significance of association with the gene set. We first applied principal component analysis (PCA) to the SNPs in each gene, and fit a logistic regression model with the resulting principal components (PCs) that explained 80% of the variation in the SNP data for each gene as the predictor variables. For each gene, this logistic regression model was used to assess the global significance of gene association by comparing the model with the PCs that explain 80% of the SNP variation in the gene with the model without the gene-specific PCs, using a likelihood ratio test (Gauderman et al., 2007). We then applied the GM to all the gene-level association p-values in a given gene set to calculate gene-set association statistics. We denote the two-step approach consisting of PCA at the gene-level followed by aggregation of gene-level p-values using GM as PC-GM. For the PC-GM approach we again used the shape parameter of 0.1 (i.e. PC-$GM_{0.1}$), but also performed a sensitivity analysis with a shape parameter of 1 (PC-GM_1). Permutations were again used to assess significance of the gene-set tests using 1000 permutations.

All analyses were limited to a subset of subjects of European-American ancestry. However, even within groups of same continental ancestry, residual population substructure can lead to false positive SNP association (Campbell et al., 2005). Moreover, because the goal of GSA is to aggregate individual SNP effects to detect the influence of a set of genes, even small population stratification effects can have substantial effects on GSA results (Fridley and Biernacka, 2011). We therefore used a PCA based approach (Price et al., 2006) to adjust for residual population structure. Three PCs were included as covariates in all analyses to adjust for effects of population stratification within the sample of European-American subjects.

Results

Table 1 shows results for pathways with an uncorrected $p<0.05$ for at least one of the primary analyses ($GM_{0.1}$ or PC-$GM_{0.1}$). Complete results for all KEGG pathways based on analyses with all the methods are shown in Supplemental Table S1. None of the results are significant after a Bonferroni correction for the number of pathways investigated. The gene set with strongest evidence for association using the two-step approach corresponds to the "synthesis and degradation of ketone bodies" pathway (PC-$GM_{0.1}$ p=0.003; PC-GM_1 p=0.0009). This pathway is also marginally significant at the uncorrected 0.05 significance level with the one-step analysis ($GM_{0.1}$ p=0.051). QQ-plots of the results of the one-step $GM_{0.1}$ and two-step PC-$GM_{0.1}$ analyses, as well as the one step GM_1 and two-step PC-$GM_{0.1}$ sensitivity analyses, are shown in Supplemental Figure 1. The plot corresponding to the PC-GM_1 analysis emphasizes this pathway. Several other pathways are nominally significant at the 0.05 significance level with both one-step and two-step analysis methods. Notably, a gene set representing neuroactive ligand-receptor interactions, one of the candidate pathways previously reported to play a role in addictive behavior (Li et al., 2008), showed the strongest evidence of association with the one-step GM analyses (p=0.008 with both $GM_{0.1}$ and GM_1).

For the top gene set (the neuroactive ligand receptor interaction pathway) yielding strongest evidence of association based on the one-step analysis (p<0.01 with both $GM_{0.1}$ and GM_1), individual SNPs that were significant at the 0.01 level, limited to the top 25 SNPs, are listed in Table 2. Individually, none of these SNPs are significantly associated with alcohol dependence after correction for multiple testing. However, these results give some indication regarding which SNPs contributed to the overall evidence for association with the gene set. In particular, many of the most strongly associated SNPs were in glutamate receptor genes, including *GRIK2* (rs1415484, rs6936552, and rs6908225) *GRM7* (rs17046239), *GRIK1* (rs2832476), and *GRM5* (rs1903851 and rs10501681). Although not in the list of top 25 results, many other glutamate receptor SNPs in this gene set had p-values < 0.01, including SNPs in *GRIA4*, *GRIA3*, *GRIN2A*, *GRIN2C*, and *GRIK4*.

For the gene set corresponding to the "synthesis and degradation of ketone bodies" pathway, which yielded the strongest evidence of association based on the two-step analysis (p<0.01 for PC-$GM_{0.1}$ and PC-GM_1), genes that were nominally significant (p<0.05) based on the PCA that makes up the first step of the two-step analysis are listed in Table 3. For this pathway, three out of the nine genes in the gene set were associated with alcohol dependence at the uncorrected 5% significance level. Of the genes in this gene set, variation in *BDH2* (3-hydroxybutyrate dehydrogenase type 2) was most strongly associated with alcohol dependence (p=0.001).

Discussion

This paper presents the first comprehensive genome-wide gene set analysis of alcohol dependence. We previously applied GSA to investigate the association of a candidate gene set, representing the NMDA-dependent AMPA trafficking cascade pathway, with alcoholism (Karpyak et al., 2011). Rather than limiting the analysis to candidate pathways, the present study investigated the association of alcohol dependence with all pathways currently represented in the KEGG database. Our results demonstrate a potential role of the "synthesis and degradation of ketone bodies" pathway (p<0.001), and provided further support for the potential involvement of the "neuroactive ligand receptor interaction" pathway in susceptibility to alcohol dependence. Other pathways with nominal evidence of association in our study may represent additional interesting signals. For example, we found nominally significant evidence of association of alcohol dependence with the gene set representing the nitrogen metabolism pathway. This is intriguing considering that this pathway was shown to be involved in response to antidepressant treatment, in a recent study that used a combined pharmacometabolomics - pharmacogenomics approach (Ji et al.).

Our top gene set association signal implicated the “Synthesis and Degradation of Ketone Bodies” pathway in alcohol dependence. Excessive alcohol consumption is known to increase ketone bodies, especially beta-hydroxybutyrate, which causes ketoacidosis (Elliott et al., 2010; Palmer, 1983). An increase in ketone bodies in the blood lowers the pH and causes dehydration because of vomiting and diuresis. Interestingly, our results demonstrate that variation in the *BDH2* gene, responsible for beta-hydroxybutyrate formation, is associated with alcohol dependence. Since accumulation of beta-hydroxybutyrate causes aversive effects of alcohol consumption, it is likely that reduced BDH2 activity is related to increased alcohol consumption. Instead, formation of acetone through acetoacetate decarboxylase might be increased in alcohol dependent patients. Conversely, since alcohol load in alcoholics showed increased beta-hydroxybutyrate and total ketone bodies levels compared to nonalcoholic control subjects (Hirsch et al., 1998), increased beta-hydroxybutyrate levels might be an indicator or marker of excessive alcohol consumption and increased BDH2 activity might account for it, which will be unraveled by further investigations.

Interestingly, one of the top gene sets identified using the one-step GSA is the “neuroactive ligand-receptor interaction” gene set. This pathway has previously been implicated in addictive behaviors and is a strong candidate based on biological knowledge (Li et al., 2008). The other four pathways reported to be associated with drug and alcohol dependence by the KARG [http://karg.cbi.pku.edu.cn/], were not significantly associated with alcohol dependence in our analyses, except for the MAPK signaling pathway for which there was nominal evidence of association based on one of the GSA methods (p=0.029 with GM_1 analysis).

The KEGG neuroactive ligand receptor interaction gene set is a very large gene set, consisting of more than 300 genes representing a variety of signaling molecules including many types of neuroreceptors. Among those are classes of neuroreceptor genes previously implicated in alcohol use disorders, such as dopamine, serotonin, gamma-aminobutyric acid (GABA), and glutamate receptors (Kohnke, 2008). Thus, the association of this gene set with alcohol dependence observed in our study is not surprising, and confirms the effectiveness of GSA in identifying pathways that play a role in complex traits. However, because this gene set is very broadly defined, it is probably over-inclusive in this context, and the finding of association between this gene set and alcohol dependence does not provide much novel information regarding specific neurobiology underlying addiction. Furthermore, GSA of large, broadly-defined gene sets may limit opportunity for uncovering positive associations as a result of inclusion of many non-informative SNPs in genes unrelated to the trait of interest. These observations emphasize the importance of well-defined pathways for successful application of GSA.

While pathway databases, such as KEGG, need to be expanded to include a broader range of pathways, analysis of custom user-specified gene sets based on prior knowledge in a given area (see e.g. (Karpyak et al., 2011)) is also an important strategy in GSA. Despite the limitations introduced by analyzing a very broadly-defined pathway, the individual SNP results for the neuroactive ligand receptor interaction pathway shown in Table 2 point towards importance of variation in genes involved in glutamate neurotransmission for alcohol dependence susceptibility. Although the statistical methods applied here have been shown to be valid (have correct type 1 errors) and have good power compared to alternative methods, new GSA methods are still being developed. Thus, there is still some uncertainty as to which methods may be optimal for analyzing data with particular properties. In this study we applied both a one-step and a two-step GSA approach. Our previous analyses of simulated data demonstrated that generally the two-step approach is more powerful than the one-step approach (Biernacka et al., 2011). However, the power of the methods depends on numerous factors including the size of the gene set and the true underlying disease model (e.g. number of SNPs in the gene set associated with the phenotype, and their distribution within the genes that belong to the gene set). Thus, in certain situations, one-step approaches may be more powerful than two-step methods. Indeed, in the analyses presented here, the one-step methods provided greater evidence for association with the neuroactive ligand receptor interactions pathway than did the two-step methods. This may occur when there are many weakly associated SNPs in a gene set, occurring in many of the genes in the gene set. In such situations, perhaps the gene-level tests that contribute to the two-step analysis may not have adequate power to detect association, while combining all of the individual SNP effects across the gene set may detect the association. Further investigation of the situations under which the different GSA methods have advantages is needed.

Limitations of the approach applied here include gaps in current knowledge about biologically relevant pathways and the corresponding genes sets. In particular, pathways for some types of biological processes are not as well documented in public databases as others. In fact, many of the top SNPs reported by Bierut et al. (2010), including those in the *PKNOX2* gene that provided the top association signal in their GWAS, were not represented in the KEGG pathways that were analyzed here. However, the *HRH1* and *GRM5* genes that were also among the top association signals reported by Bierut et al., as well as the candidate *GABRA2* gene for which they reported nominally significant findings, are all part of the neuroactive ligand receptor interactions pathway that we identified in our analyses.

Another limitation arises from the fact that some genes/pathways may not be as well represented with the available genome-wide SNP arrays as other genes/pathways. Thus, power to detect associations with different pathways is not expected to be uniform. Finally, the optimal way of assigning SNPs to pathways is still not clear. We included all genotyped SNPs within 5kb of the gene (i.e. defined from first to last exon). Although this is consistent with what is usually done in GSA (Fridley and Biernacka, 2011), it is an arbitrary distance. Determining which SNPs should be included in a gene set based on other criteria such as eQTLs (SNPs that are associated with the expression level of a gene, either through cis or trans effects), has been proposed (Zhong et al., 2010). However, the utility of this approach has not been thoroughly investigated, and data on relevant expression (e.g. brain) eQTLs are currently limited. Therefore at this point a simple distance-based SNP inclusion criterion was used. However, future studies should consider other ways of defining gene sets. Future research should also include investigation of gene sets defined using other pathway databases (e.g. Gene Ontology or MetaCore), as well as application of novel methods, for example approaches that take into account gene-gene interactions.

It is important to note that the GSA applied in this study does not attempt to identify individual loci associated with the phenotype, and therefore does not identify specific SNPs for functional study. Nevertheless, the approach identifies sets of genes representing relevant pathways, enabling further more focused biomarker studies of complex traits.

In summary, analyses of individual SNPs from the SAGE data did not identify any genome-wide significant results, and the top results did not replicate in two independent replication samples included in the study of Bierut et al. (Bierut et al., 2010). Here, by using a novel approach that looks at the collective evidence of association with a set of SNPs in related genes, new leads for further investigation have been identified. This study demonstrates the utility of GSA in the analysis of complex disease data, and suggests specific directions for further research into the genetic architecture of alcohol dependence. Independent replication, functional validation, and more in-depth analyses, such as investigation of gene-gene interactions, are warranted for the top pathways, including “synthesis and degradation of ketone bodies” and “neuroactive ligand receptor interactions”.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by a grant from the National Institute on Alcohol Abuse and Alcoholism (NIAAA; R03 AA019570). BLF was supported by grants from NIGMS and NCI (R21GM86689 and R21 CA140879). The funding agencies had no role in study design, analysis, interpretation of results, decision to publish, or preparation of the manuscript. We also acknowledge the Study of Addiction: Genetics and Environment (SAGE), which is part of the Gene Environment Association Studies (GENEVA) initiative supported by the National Human Genome Research Institute (dbGaP study accession phs000092.v1.p1).

References

Agrawal A, Lynskey MT. Are there genetic influences on addiction: evidence from family, adoption and twin studies. Addiction. 2008; 103(7):1069–1081. [PubMed: 18494843]

Ball D. Addiction science and its genetics. Addiction. 2008; 103(3):360–367. [PubMed: 18042191]

Biernacka JM, Jenkins GD, Wang L, Moyer AM, et al. Use of the gamma method for self-contained gene-set analysis of SNP data. European Journal of Human Genetics. 2012; 20:565–571. [PubMed: 22166939]

Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, et al. A genome-wide association study of alcohol dependence. Proceedings of the National Academy of Sciences U S A. 2010; 107(11):5082–5087.

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, et al. Demonstrating stratification in a European American population. Nature Genetics. 2005; 37(8):868–872. [PubMed: 16041375]

Chai HS, Sicotte H, Bailey KR, Turner ST, et al. GLOSSI: a method to assess the association of genetic loci-sets with complex diseases. BMC Bioinformatics. 2009; 10:102. [PubMed: 19344520]

Elliott S, Smith C, Cassidy D. The post-mortem relationship between beta-hydroxybutyrate (BHB), acetone and ethanol in ketoacidosis. Forensic Science International. 2010; 198(1-3):53–57. [PubMed: 19954904]

Fisher, R. Statsitical Methods for Research Workers. London: Oliver and Boyd; 1932.

Fridley BL, Biernacka JM. Gene set analysis of SNP data: benefits, challenges, and future directions. European Journal of Human Genetics. 2011; 19(8):837–843. [PubMed: 21487444]

Fridley BL, Jenkins GD, Biernacka JM. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. PLoS One. 2010; 5(9):e12693. [PubMed: 20862301]

Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genetic Epidemiology. 2007; 31(5):383–395. [PubMed: 17410554]

Goldman D, Oroszi G, Ducci F. The genetics of addictions: uncovering the genes. Nature Reviews Genetics. 2005; 6(7):521–532.

Hirsch S, De La Maza MP, Petermann M, Bunout D. Lipid turnover in alcoholics before and after an ethanol load. Nutrition. 1998; 14(5):437–442. [PubMed: 9614308]

Holmans P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. Advances in Genetics. 2010; 72:141–179. [PubMed: 21029852]

Holmans P, Green EK, Pahwa JS, Ferreira MA, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. American Journal of Human Genetics. 2009; 85(1):13–24. [PubMed: 19539887]

Ji Y, Hebbring S, Zhu H, Jenkins GD, et al. Glycine and a glycine dehydrogenase (GLDC) SNP as citalopram/escitalopram response biomarkers in depression: pharmacometabolomics-informed pharmacogenomics. Clinical Pharmacology and Therapeutics. 2011; 89(1):97–104. [PubMed: 21107318]

Karpyak VM, Geske JR, Colby CL, Mrazek DA, et al. Genetic variability in the NMDA-dependent AMPA trafficking cascade is associated with alcohol dependence. Addiction Biology. 2012; 17:798–806. [PubMed: 21762291]

Kohnke MD. Approach to the genetics of alcoholism: a review based on pathophysiology. Biochemical Pharmacology. 2008; 75(1):160–177. [PubMed: 17669369]

Kronenberg F. Genome-wide association studies in aging-related processes such as diabetes mellitus, atherosclerosis and cancer. Experimental Gerontology. 2008; 43(1):39–43. [PubMed: 17967522]

Lettre G, Rioux JD. Autoimmune diseases: insights from genome-wide association studies. Human Molecular Genetics. 2008; 17(R2):R116–121. [PubMed: 18852199]

Li CY, Mao X, Wei L. Genes and (common) pathways underlying drug addiction. PLoS Computational Biology. 2008; 4(1):e2. [PubMed: 18179280]

O'Dushlaine C, Kenny E, Heron E, Donohoe G, et al. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Molecular Psychiatry. 2011; 16:286–292. [PubMed: 20157312]

Palmer JP. Alcoholic ketoacidosis: clinical and laboratory presentation, pathophysiology and treatment. Clinical Endocrinology and Metabolism. 1983; 12(2):381–389.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38(8):904–909. [PubMed: 16862161]

Ruano D, Abecasis GR, Glaser B, Lips ES, et al. Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. American Journal of Human Genetics. 2010; 86(2):113–125. [PubMed: 20060087]

Srinivasan BS, Doostzadeh J, Absalan F, Mohandessi S, et al. Whole genome survey of coding SNPs reveals a reproducible pathway determinant of Parkinson disease. Human Mutation. 2009; 30(2): 228–238. [PubMed: 18853455]

Treutlein J, Cichon S, Ridinger M, Wodarz N, et al. Genome-wide association study of alcohol dependence. Archives of General Psychiatry. 2009; 66(7):773–784. [PubMed: 19581569]

Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. American Journal of Human Genetics. 2007; 81(6)

Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics. 2010; 11(12):843–854.

Zaykin DV, Zhivotovsky LA, Czika W, Shao S, et al. Combining p-values in large-scale genomics experiments. Pharmaceutical Statistics. 2007; 6(3):217–226. [PubMed: 17879330]

Zhong H, Yang X, Kaplan LM, Molony C, et al. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. American Journal of Human Genetics. 2010; 86(4):581–591. [PubMed: 20346437]

Table 1

Top results of GSA. Pathways are ordered based on the 2-step PC-$GM_{0.1}$ GSA p-value. Only pathways with p<0.01 based on at least one of the primary analysis methods are shown. Results for all pathways are shown in Supplemental Table S1.

KEGG pathway ID	KEGG pathway name	N genes	N SNPs	1-step $GM_{0.1}$ GSA p-value	2-step PC-$GM_{0.1}$ GSA p-value
72	Synthesis and degradation of ketone bodies	9	128	0.051	0.003
3320	PPAR signaling pathway	69	1377	0.042	0.004
360	Phenylalanine metabolism	22	337	0.004	0.235
4080	Neuroactive ligand-receptor interaction	302	8682	0.008	0.073
4110	Cell cycle	128	2154	0.010	0.11

[a] For pathway 72, 10,000 permutations were used to obtain the 2-step GSA p-value

Table 2

Top SNP-specific association test results for the "Neuroactive ligand-receptor interaction" pathway (KEGG id: 4080), which was identified using the 1-step analysis. The top 25 SNPs with smallest p-values are shown.

SNP	gene	p
rs433303	HRH1	0.00012
rs2072100	TAC1	0.00021
rs1229434	TAC1	0.00028
rs1415484	GRIK2	0.00035
rs1848845	TAC1	0.00048
rs874306	GLP2R	0.00059
rs6936552	GRIK2	0.00068
rs17046239	GRM7	0.00072
rs2832476	GRIK1	0.00081
rs1903851	GRM5	0.00119
rs237899	OXTR	0.00123
rs443137	HRH1	0.00123
rs7756097	HCRTR2	0.00137
rs430353	HRH1	0.00139
rs9822871	HRH1	0.00140
rs17681708	GLP2R	0.00158
rs7916403	HTR7	0.00175
rs709024	ADRA1D	0.00188
rs10501681	GRM5	0.00194
rs17681684	GLP2R	0.00206
rs12736154	LEPR	0.00206
rs12345664	GABBR2	0.00208
rs17676067	GLP2R	0.00208
rs995213	GABBR2	0.00210
rs6908225	GRIK2	0.00239

The total number of analyzed SNPs in this pathway was 8682.

Gene name abbreviations: HRH1 = histamine receptor H1; TAC1 = tachykinin, precursor 1; GRIK2,1 = glutamate receptor, ionotropic, kainate 2,1; GLP2R = glucagon-like peptide 2 receptor; GRM7,5 = glutamate receptor, metabotropic 7,5; OXTR = oxytocin receptor; HCRTR2 = hypocretin (orexin) receptor 2; HTR7 = 5-hydroxytryptamine (serotonin) receptor 7 (adenylate cyclase-coupled); ADRA1D = adrenergic, alpha-1D-, receptor; LEPR = leptin receptor; GABBR2 = gamma-aminobutyric acid (GABA) B receptor, 2.

Table 3

Top gene-level association results in the top pathway (Synthesis and degradation of ketone bodies; KEGG ID 72) selected based on 2-step analysis. Genes with p-value < 0.05 are listed.

gene	Chromosome location	P
BDH2 (3-hydroxybutyrate dehydrogenase, type 2)	4q24	0.001
OXCT1 (3-oxoacid CoA transferase 1)	5p13.1	0.021
ACAT1 (acetyl-CoA acetyltransferase 1)	11q22.3	0.040

The total number of genes in this gene set was 9.