Genetic-Demographic Dissection of Alcoholism

Amin Zollanvari, PhD[1,2*]; James Thomas[2*]; Gil Alterovitz, PhD[1,2,3]

[1]Center for Biomedical Informatics, Harvard Medical School, Boston, MA; [2] Children's Hospital Informatics Program at Harvard-MIT Division of Health Science, Boston, MA; [3] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA

*** These authors contributed equally to this work.**

Abstract

We assessed 3,776 individuals to construct a model capable of using demographic, clinically ascertainable, and genetic factors for predicting alcohol dependence in individuals. Our Bayesian network-based framework provides the first predictive model of alcohol dependence. The network has a predictive power of 92%, as measured by the area under the receiver operating characteristic curve (AUROC). This is thus the first "excellent" predictor (per on objective metrics) of alcoholism. It achieved significant improvement over models that we designed using only genetic variables (AUROC of 81%), demographic variables (AUROC of 80%), or demographic variables and previously used clinically ascertainable metrics (AUROC of 80%). We demonstrate that the improvement was not simply due to an increase in the number of variables, but rather to a synergy between two types of variables: demographics and genetics. Furthermore, through analysis of our predictive networks, we identify novel genes and 14 biological pathways that are likely to be associated with the development of alcoholism. Our approach can be used in the future to construct and analyze a model of others disease for which genetic and demographic case-control data is available.

1. Introduction

Alcohol dependence is characterized by increasing tolerance to and consumption of alcohol, even in the face of adverse effects [1]. Almost 14% of alcohol consumers in the United States meet the criteria for alcohol dependence at some point in their lifetimes [2]. The consequences of alcohol dependence are severe. Overconsumption of alcohol is known to be a contributing factor to more than 60 diseases, including several types of cancer, and accounts for approximately 2.5 million deaths each year [3].

Alcoholism is very difficult to overcome once it initiates, and thus there has been much interest in preventing the onset of alcoholism altogether [3]. While a number of factors have been associated with alcoholism, there is a need for accurate predictive models that integrate heterogeneous information. Such a model would enable preventative intervention in cases where the onset of alcoholism appears likely as well as help in understanding the underlying biological mechanisms and risk factors. The construction of a genetic model of alcoholism has become increasingly possible with new genetic case-control studies of the disease [2]. Indeed, alcoholism is particularly amenable to a genetic model, as the genetic basis of the disease is strong. Adoption studies have demonstrated that children with alcoholic biological parents are likely to become alcoholics themselves, even if they are reared by adoptive parents in environments with few traces of alcohol [4]. Most adoption and twin studies suggest that 50-80% of variation in the phenotype is due to genetic factors [5, 6]. That different people have different initial levels of tolerance to alcohol and thus different propensities to become physically addicted to it is further evidence of the genetic basis of the disease. That said, the same studies that have pointed to genetic factors have shown that demographic factors such as culture and level of education also contribute to alcoholism [7]. Thus, an effective model of alcoholism should incorporate both demographic and genetic information.

There have been several association studies that have sought to identify a small number of susceptibility loci for alcoholism [3,8]. However, complex traits like alcoholism are commonly underpinned by *numerous* factors, genetic as well as demographic, each of which has a small effect size [9]. Thus, many genome-wide association (GWA) studies on alcoholism have struggled to pinpoint individual single nucleotide polymorphisms (SNPs) that explain a good portion of the variation in the phenotype; the best odds ratios for individual SNPs reported in a recent study were around 2 [2], a relatively low figure. Rather than association as in typical GWA analyses, what is needed is high *predictive* power. In this work, we designed a Bayesian network-based model with this goal in mind by capturing various factors: from SNPs to demographic variables as well as their interactions. We analyze the model to determine the demographic factors, genes, biological pathways, and interactions among factors that are most related to alcoholism and compare our findings to the literature.

2. Results and Discussion

2.1 Combination of Demographic and Genetic Features

The cost of genotyping an individual has been falling quickly [30]. Thus, use of genetic information is becoming feasible at a large scale. We examined two types of easily discernible variables in this study: those that we called demographic (Table 1) and those that are clinically ascertainable via family history of alcoholism. With an AUROC of 92%, the best constructed demographic-genetic model presented a substantial improvement over the models based on demographic variables alone (AUROC of 80%), family history variables alone (AUROC of 58%), and a combination of demographic and family history variables (AUROC of 80%). Although one might expect a small improvement to occur when the two sets of features (genetic and demographic) are combined, an increase in predictive power of more than 12% indicates a strong synergy between the two sets of features. The existence of this synergy was further confirmed when we created another demographic-genetic classifier with the same number of features as the best genetic-only classifier with 363 Single Nucleotide Polymorphisms (SNPs) (see Methods section 3.4). This demographic-genetic classifier had an AUROC of 91%, (rated as an "excellent" predictor) as opposed to the 81% AUROC (rated as a "good" predictor) of the genetic-only classifier.

With an AUROC of 92% ($p<0.0001$), our demographic-genetic classifier, which has 428 features, is considered an "excellent" predictor based on objective metrics [29]. Figure 1 provides the full picture of the demographic-genetic Bayesian network. Twenty-five of the 413 SNPs in the demographic-genetic classifier are found in the 21 loci previously linked to alcohol dependence ($p<0.0025$), providing evidence that our model is meaningful based on known biology. Figure 2 is a sub-graph of the network presented in Figure 1.

The predictive power of our demographic-genetic classifier confirms the frequent assertion that alcoholism is a byproduct of genetic and demographic factors. The classifier provides, for the first time, a means to accurately determine a person's risk for alcohol dependence given demographic and genetic information. Based on Figure 2, there seem to be a few likely reasons why such a synergy exists between demographic and genetic variables. First, the inclusion of race allowed the Bayesian network to distinguish between SNPs that increase the risk of alcoholism only in African Americans (AAs) and those that do so only in European Americans (EAs). It is clear from Figure 2 that there are a large number of SNPs that fit that description. As further evidence of race's role, removal of race from the demographic-genetic classifier results in a decline in AUROC of 8.7%, the largest decline of any feature, but the AUROC of a classifier with race as its only feature is 54%, which is only 4% better than a coin flip. This suggests that the race variable derives its predictive power mostly from associations with other features, namely SNPs.

2.2 Analysis of Network Composition

2.2.1 Prediction-based Analysis of the Constructed Network

We sought to determine the most important genes, demographic features, biological pathways, and feature-feature relationships represented in our genetic-demographic predictive model, henceforth called $C_{dem\text{-}gen}$. For the genes, we enumerated all genes with at least one SNP in $C_{dem\text{-}gen}$. For each such gene, we constructed and recorded the AUROC of a classifier with a feature set of the SNPs in $C_{dem\text{-}gen}$ within that gene, as well as race and sex, which would unlock the full potential of race- or sex-specific SNPs. We next considered important demographic features. It appeared that many demographic features (especially race and sex) derived most of their predictive power from edges with SNPs. Therefore, to evaluate the importance of each demographic feature, we calculated the decline in resubstitution AUROC (AUROC on the training set) upon removal of that feature and all edges connected to it from $C_{dem\text{-}gen}$. Resubstitutuion was used because the response of cross-validation AUROC to minor changes in a classifier is relatively imprecise due to larger variance of cross-validation estimators [45]. We used the Molecular Signatures Database [27] to determine the lists of genes related to 186 pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [28]. For each KEGG pathway, we recorded the AUROC of a classifier with a feature set of just the SNPs in $C_{dem\text{-}gen}$ within the pathways's genes, as well as race and sex. Finally, for the feature-feature relationships, we successively removed each edge in $C_{dem\text{-}gen}$ between non-phenotype features and recorded the decline in AUROC. Now, we were left with an AUROC for each gene and pathway, and a decline in AUROC for each demographic feature and feature-feature relationship. Then we determined the statistical significance of each gene or pathway AUROC (see Methods section 3.4) as well as the statistical significance of decline in AUROC for each demographic variable.

Our analysis, the results of which are shown in Tables 3-6, sought to dissect our best classifier, $C_{dem\text{-}gen}$, to identify strong associations between alcoholism and genes, demographic variables, biological pathways, and interactions among factors. The literature explicitly confirmed some of the identified associations, providing further evidence that our classifier is not spurious. In other cases, we found evidence in the literature suggestive of the validity of associations. A few associations were not corroborated in the literature, but the general alignment of our results with prior work suggests that insight into the emergence of alcoholism. These associations are worthy candidates for further study.

2.2.2 Genes

Alcohol has a variety of effects on the body; many of these arise from alcohol's activation of receptors in the brain [31]. A number of the genes identified in our analysis have important functions in the brain. In total, 9 of the 13 genes listed in Table 4 (excluding the intergenic set) either have been explicitly associated with alcoholism in the literature or have functional ties to the disease (e.g. are involved in brain activity). Three genes have been explicitly associated with the development of alcoholism. *CPE* has been identified in prior GWA studies on alcoholism [8], and it encodes the enzyme carboxypeptidase E, which activates neuropeptides [32], proteins crucial to communication among neurons. *PKNOX2*, which regulates the transcription of other genes and affects anatomical development [33], has been linked to all types of substance abuse in European women [34]. *GLT25D2* was identified as related to alcoholism in a GWA study on a dataset that had no samples in common with ours [35]. Five other genes have functional ties to alcoholism and the development of the behavior (Supplementary Information). While many identified genes were generally in alignment with prior knowledge, further work should be done to understand the associations between alcoholism and the five genes that went uncorroborated in the literature (*BLNK*, *BMPER*, *PDLIM5*, *VEPH1*, *AMPD3*). Finally, the high importance of intergenic SNPs is surprising, but similar SNPs have been tied in prior GWA studies to alcoholism [18], and the noncoding RNA that is transcribed from intergenic regions affects gene expression levels in some cases [36]. Several other genes with two or more SNPs are involved in the predictive model (Supplementary Information).

2.2.3 Demographic Features and Feature-Feature Relationships

Table 5 and 6 present the demographic variables and the links between features that are "significant" predictors of phenotype (see . Some of these variables and links are explicitly stated in prior studies (for details see section 2.2.1 and Methods section 3.4). Thus, they serve primarily to show the ability of our model in unifying the effect of known factors. For example, a prior study [37] has demonstrated that alcohol consumption is negatively correlated with both income and educational status, both of which were deemed significant demographic features. The significance of the edge between income and education is sensible as well, as the conditional probability tables of the classifier indicate that a high level of education may be able to counteract a low level of income with respect to the development of alcoholism, and vice versa. Another prior study [38] provides the reason for the significance of the edge between race and income: there is a much stronger association between income and alcoholism in African Americans than in European Americans. Although no SNP-SNP edges were deemed significant, the numerous SNP-SNP edges that connect SNPs on the same gene (see Figure 2) are reasonable, as SNPs that are closer together are more likely to interact and/or affect the same function [10].

There is also a significant predictive link between race and rs8225 in Table 6 (decline in AUROC has $p<0.04$). While we used a prediction-based approach to identify this significant link, one may realize the importance of such link by examining the distribution of rs8225 among cases and controls in both races (see Table 7). As presented in Table 7, the distribution of this variant is substantially different between the two race groups in both cases and controls (difference of distribution of AAs and EAs in controls has a $p<10^{-15}$

and in cases $p<10^{-15}$ as determined by Cochran-Armitage test [46]). The within race group distribution of this variant is also significantly different between cases and controls (difference of distribution of AAs in controls and cases has a $p<0.005$ and this difference for EAs has $p<0.0002$ as determined by Cochran-Armitage test [46]). There (decline in AUROC has p<0.02); rs5933820 is located on the X chromosome.

2.2.4 Biological Pathways.

Twelve of the 14 biological pathways discovered in our analysis have already been linked, either explicitly or indirectly, to alcoholism in the literature. (Italicized terms in this section correspond to the biological pathways listed in Table 3.) Two pathways have been explicitly cited for their involvement in the development of alcohol dependence. For example, Fombonne, *et al.* demonstrated that children with *long-term depression* are at higher risk for alcohol dependence in adulthood [39]. The binding of GABA receptors, which are *neuroactive ligand receptors*, was found to be abnormally high in the brains of alcoholics [40]. Evidence in the literature suggests that four pathways may be involved in the emergence of alcoholism. It has been noted that alcohol inhibits the reorganization of the *actin cytoskeleton* [41]. Chronic exposure to alcohol reduces *calcium signaling* in response to glutamate receptor stimulation in neuronal cells [42]. Exposure of intestinal Gram negative bacteria to alcohol results in accumulation of acetaldehyde, which in turn increases tyrosine phosphorylation of *adherens junction* proteins [43]. Treatment of the ventral tegmental area in mice with glial cell line-derived neurotrophic factor activated the *MAPK signaling pathway* and reduced desire for alcohol [44]. Six pathways do not seem likely to be involved in the onset of alcoholism, but do appear to have links to the behavior (for more information in this regard see Supplementary Information). Due to the overall alignment of the results of the analysis with the literature, it is likely that the two pathways that have not yet been explicitly tied in some way to alcoholism (dilated cardiomyopathy and hypertrophic cardiomyopathy) have links to the behavior; further study is required to confirm such links.

3. Materials and Methods

3.1 Bayesian Approach to Dissecting Alcoholism

A Bayesian network is a directed acyclic graph that compactly represents the joint probability distribution of a set of variables. Bayesian networks have previously proven effective on this front: they have successfully described the complex interactions underpinning polygenic traits such as early stroke under the context of sickle cell anemia [10] and nicotine dependence [11]. In a Bayesian network, nodes represent variables and edges represent probabilistic dependencies between variables. Since Bayesian networks represent joint distributions, they can be used to predict the probability of observing a specific state of a target variable (in our case, a phenotype) given the states of all other variables (in our case, SNPs and demographic variables), and have consequently been used as classifiers.

In this work, we employ a maximal weighted spanning tree approach to infer Bayesian network structure [12]. A Bayesian network constructed with this approach is called a tree-augmented naive Bayesian network (TAN) [13]. A TAN has a relatively simple structure, as each non-class (in our case, non-phenotype) variable has edges only to the class and the one other variable with which it has highest mutual information. It has been shown that such a classifier commonly outperform many commonly used classifiers [13, 14] including naive Bayesian networks, which are incapable of inferring the pattern of dependencies among variants. For these reasons, constructing a predictive model as described above allows us to capture the effects of many genetic and demographic variables on alcoholism.

3.2 Data Collection and Pruning

We utilized SAGE data [2], which featured 3,829 subjects and considered 948,658 SNPs from across the human genome, as well as several demographic variables. The data included human samples from three prior studies [2]; 30% of the individuals were African Americans and 70% were European Americans. The SAGE dataset includes 1,897 *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* cases and 1,932 alcohol-exposed non-dependents. We first removed any SNPs out of Hardy-Weinberg equilibrium ($P < 0.0001$). Hardy-Weinberg equilibrium tests were run separately on the African Americans and the European Americans in order to ensure identification of any SNPs common only in one race out of equilibrium. SNPs with minor allele frequency (MAF) below 0.01 or call rate below 98% were also removed from consideration, leaving a total of 934,128 SNPs. Finally, the 3,776 samples with a genotyping rate above 98% were maintained.

3.3 Model Construction

In addition to our demographic-genetic model, we constructed four other Bayesian models, which served as benchmarks for evaluating our demographic-genetic model. The first benchmark model included only genetic factors, the second only demographic factors, the third both demographic factors and information on family history of alcoholism, and the last only information on family history. The demographic-only and genetic-only models enabled us to quantify the improvement provided by combining genetic and

demographic factors. We constructed the models with family history information, which is very easy to collect, to see whether the predictive power of such information compares to that of genotypic data, which is more difficult to obtain. Depending on how many SNPs contribute to alcoholism and the dominance patterns of such SNPs, family history variables should capture some amount of the information content from the genotypic data. All of the models were trained with the 3,776 remaining SAGE samples. All areas under the receiver operating characteristic curve (AUROCs) for the models were determined by 3-fold cross-validation.

Demographic-Genetic Model. We first constructed our demographic-genetic model. A Cochran-Mantel-Haenszel (CMH) association test was used to rank the 934,128 SNPs [46]. The association analysis was performed with the software PLINK [15]. Only the 654 SNPs ($p < 0.0005$) were maintained. We used 15 demographic variables, listed in Table 1. Several demographic variables were left out, especially ones relating to comorbidities, because their distributions across the cases and controls were heavily imbalanced. All continuous demographic variables (e.g. income) were discretized, as the software we used to construct Bayesian networks required discrete features. To produce an optimal classifier, two approaches were employed: an iterative approach and one based on a combination of linkage disequilibrium (LD) analysis [16] and the iterative approach. In the iterative approach, a Bayesian network (Section 3.1) was first trained with the remaining SNPs and the 15 demographic variables as features. In each subsequent iteration, the 50 SNPs with the highest p-values were removed from the feature set and a new classifier was trained. The best classifier was the one with the highest AUROC. The LD analysis-based approach sought to eliminate redundant SNPs. LD analysis was performed and SNPs that were strongly linked (i.e. frequently co-occurred in both the cases and the controls) were grouped into bins. The approach outlined by Carlson, *et al.* [17], with the r^2 threshold lowered from 0.8 to 0.4, was used to produce a single tag SNP for each LD bin. Only the tag SNPs were maintained, and the iterative approach was applied to them. This approach ensures that multiple SNPs that are proxies due to low LD distance are not selected. The tag SNP acts as a proxy for all SNPs in that region. The best classifiers from the two approaches were compared, and the one with the highest AUROC was selected as the optimal demographic-genetic classifier.

Demographic-Only Model. To construct the demographic-only model, only the 15 variables listed in Table 1 were used as features. Continuous variables were again discretized.

Genetic-Only Model. The initial feature set for the genetic-only model consisted of 654 SNPs ($p < 0.0005$). The iterative and the LD analysis-based approaches were used to cull the feature set, and the constructed classifier with the highest AUROC was selected as the best genetic-only model.

Demographic-Family History Model. The feature set for this model consisted of the 15 demographic variables as well as two binary variables provided by the dataset on whether or not a subject's mother and father had alcohol dependence. Continuous variables were discretized as before.

Family History-Only Model. Only the two aforementioned family history variables were used as features in this classifier.

3.4 Analysis of Demographic-Genetic Model

We undertook a number of efforts to analyze the quality and structure of our demographic-genetic prognostic model ($C_{dem\text{-}gen}$).

Verification of Demographic-Genetic Synergy. We sought to verify that any improvement in predictive power upon incorporation of both demographic and genetic variables into a model was not simply attributable to an increase in the number of features, but rather to a synergy between the two types of variables. Thus, we took the feature set used in the best classifier from the first analysis (genetic-only), replaced the 15 SNPs with highest p-values with the 15 demographic variables, and created a new classifier with the modified (but with size unchanged) feature set. If the AUROC of the new classifier was higher than that of the best genetic-only classifier, the combination of demographic and genetic factors was deemed to have a synergistic effect.

Statistical Significance of AUROC of the Prognostic model. To verify that overfitting was not responsible for our results, we determined the statistical significance of the AUROC of $C_{dem\text{-}gen}$. We permuted the labels to create thousand random datasets. Therefore, each random dataset was constructed by assigning a random phenotype to each sample in the original SAGE dataset. All of the steps described in Section 3.2 were carried out on each random dataset, as were the steps in the first subsection of Section 3.3. The thousand AUROCs for the thousand resultant classifiers were used to determine a p-value for the AUROC.

Alignment with Previously Identified Susceptibility Loci. The 21 chromosomal regions listed in Table 2 have been identified as related to alcoholism in previous association or linkage studies (all of which employed datasets and/or statistical methods different from ours) [2, 18–26]. By number of base pairs, the 21 regions make up approximately 3% of the human genome. We determined the number of SNPs in $C_{dem\text{-}gen}$ that are found in the regions. We calculated the p-value of this number under the null hypothesis that the appearance of SNPs from the regions in the model was completely random.

Statistical Significance of AUROC of Each Gene, Pathway, or Demographic Variables. To determine the statistical significance of each gene or pathway AUROC, we constructed 1,000 classifiers, each with the same number of features as the classifier for which the AUROC in question was calculated, and determined their AUROCs. The set of genetic features for each of the 1,000 classifiers was drawn randomly from the background set of SNPs. Race and sex were included as features in all 1,000 classifiers in order to ensure parity with the procedure used to generate AUROCs for genes and pathways. The list of 1,000 random AUROCs enabled the calculation of a p-value for the AUROC in question.

To determine the statistical significance of each decline in AUROC (used for quantifying the significance of predictive power of each demographic variable and link between features), we used the same background set to construct 1,000 random classifiers with the same set of demographic variables and the same number of genetic features as $C_{dem\text{-}gen}$. For each randomly generated model, we recorded the decline in AUROC upon removal of a random genetic feature (in the case of the declines in AUROC for demographic features) or a random edge (in the case of the declines in AUROC for feature-feature relationships). The 1,000 random declines in AUROC enabled the calculation of a p-value for the decline in AUROC of interest. Each gene, demographic feature, pathway, and feature-feature relationship was now associated with a p-value. A p-value of 0.05 was used as the threshold for statistical significance.

5. Acknowledgements

We thank Kent Huynh for his contribution in implementing the pipeline, and Aaron Merlob for critically reviewing the manuscript. This work was supported by grants 5R21DA025168-02 (G. Alterovitz), 1R01HG004836-01(G. Alterovitz), and 4R00LM009826¬03 (G. Alterovitz).

6. Author Contributions

A.Z. provided the bioinformatics background, designed the study, and helped draft the manuscript. J. T. implemented the pipeline, participated in the design, and drafted the manuscript. G.A. participated in the design and coordination of the study and helped draft the manuscript.

7. Conflict of Interest

The authors declare that they have no conflict of interest.

8. References

[1] Li, T. K., *et al.* The Alcohol Dependence Syndrome, 30 years later: a commentary. *Addiction* 102 (2007): 1522-30.
[2] Bierut, L. J., *et al.* A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci USA* 107 (2010): 5082-7.
[3] World Health Organization. Global Status Report on Alcohol and Health 2011. Geneva, Switzerland. 2011.
[4] Agrawal, A., *et al.* Are there genetic infuences on addiction: evidence from family, adoption and twin studies. *Addiction* 103 (2008): 1069-1081.
[5] Heath, A. C., *et al.* Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. *Psychol Med* 27 (1997): 1381-96.
[6] Knopik, V. S., *et al.* Genetic effects on alcohol dependence risk: re-evaluating the importance of psychiatric and other heritable risk factors. *Psychol Med* 34 (2004): 1519-30.
[7] Prescott, C. A., *et al.* Genetic and Environmental Contributions to Alcohol Abuse and Dependence in a Population-Based Sample of Male Twins. *Am J Psychiatry* 156 (1999): 34-40.
[8] Edenberg, H. J., *et al.* Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcohol Clin Exp Res* 34 (2010): 840-852.
[9] Zondervan, K. T., *et al.* The complex interplay among factors that influence allelic association. *Nature Reviews Genetics* 5 (2004): 89-100.
[10] Sebastiani, P., *et al.* Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nature Genetics* 37 (2005): 435-40.
[11] Ramoni, R. B., *et al.* A testable prognostic model of nicotine dependence. *J Neurogenet* 23 (2009): 283-92.
[12] Chow, C. K., *et al.* Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory* IT-14 (1968): 462-467.
[13] Friedman, N., *et al.* Bayesian Network Classifiers. *Machine Learning* 29 (1997): 131-163.
[14] Cheng, J., *et al.* Comparing Bayesian Network Classifiers. *Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence* (1999): 101-108.
[15] Reich, D. E., *et al.* Linkage disequilibrium in the human genome. *Nature* 411 (2001): 199-204.
[16] Purcell, S., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81 (2007): 559-75.
[17] Carlson, C. S., *et al.* Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. *American Journal of Human Genetics* 74 (2004): 106-120.
[18] Treutlein, J., *et al.* Genome-wide association study of alcohol dependence. *Arch Gen Psychiatry* 66 (2009): 773-84.
[19] Long, J. C., *et al.* Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population. *Am J Med Genet* 81 (1998): 216-21.
[20] Reich, T., *et al.* Genome-wide search for genes affecting the risk for alcohol dependence. *Am J Med Genet* 81 (1998): 207-15.

[21] Thomasson, H. R., *et al.* Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. *Am J Hum Genet* 48 (1991): 677-81.
[22] Saccone, N. L., *et al.* A genome screen of maximum number of drinks as an alcoholism phenotype. *Am J Med Genet* 96 (2000): 632-7.
[23] Dick, D. M., *et al.* The genetics of alcohol and other drug dependence. *Alcohol Res Health* 31 (2008): 111-8.
[24] Edenberg, H. J., *et al.* The genetics of alcoholism: identifying specific genes through family studies. *Addict Biol* 11 (2006): 386-96.
[25] Dick, D. M., *et al.* Candidate genes for alcohol dependence: a review of genetic evidence from human studies. *Alcohol Clin Exp Res* 27 (2003): 868-79.
[26] Wang, K. S., *et al.* A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence. *J Psychiatr Res* (2011). doi:10.1016/j.jpsychires.2011.06.005.
[27] Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102 (2005): 15545-50.
[28] Kanehisa, M., *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28 (2000): 27-30.
[29] Pines, J. M., and W. W. Everett. *Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules*. BMJ Books, 2008.
[30] Ho, L. A., *et al.* Using public control genotype data to increase power and decrease cost of case-control genetic association studies. *Hum Genet* 128 (2010): 597-608.
[31] Platt, D. M., *et al.* Contribution of a1GABAA and a5GABAA Receptor Subtypes to the Discriminative Stimulus Effects of Ethanol in Squirrel Monkeys. *Journal of Pharmacology and Experimental Therapeutics* 313 (2005): 658-667.
[32] Hook, V., *et al.* Proteases for processing proneuropeptides into peptide neurotransmitters and hormones. *Annu Rev Pharmacol Toxicol* 48 (2008): 393-423.
[33] Imoto, I., *et al.* Identification and Characterization of Human PKNOX2, a Novel Homeobox-Containing Gene. *Biochemical and Biophysical Research Communications* 287 (2001): 270-276.
[34] Chen, X., *et al.* PKNOX2 gene is significantly associated with substance dependence in European-origin women. *Proc Natl Acad Sci USA* 106 (2009): 17241.
[35] Agrawal, A., *et al.* A Candidate Gene Association Study of Alcohol Consumption in Young Women. *Journal of Studies on Alcohol and Drugs* 35 (2011): 550-558.
[36] Rusk, N. Noncoding transcripts as expression boosters. *Nature Methods* 7 (2010): 947.
[37] Midanik, L. T., *et al.* The Demographic Distribution of US Drinking Patterns in 1990: Description and Trends from 1984. *American Journal of Public Health* 84 (1994): 1218-22.
[38] Barr, K., *et al.* Race, Class, and Gender Differences in Substance Abuse: Evidence of Middle-Class/Underclass Polarization among Black Males. *Social Problems* 40 (1993): 314-327.
[39] Fombonne, E., *et al.* The Maudsley long-term follow-up of child and adolescent depression. *Br J Psychiatry* 179 (2001): 218-23.
[40] Tran, V. T., *et al.* GABA receptors are increased in brains of alcoholics. *Annals of Neurology* 9 (1981): 289-292.
[41] Dai, Q., *et al.* Ethanol Suppresses LPS-induced Toll-like Receptor 4 Clustering, Reorganization of the Actin Cytoskeleton, and Associated TNF-a Production. *Alcohol Clin Exp Res* 30 (2006): 1436-44.
[42] Gruol, D. L., *et al.* Chronic alcohol reduces calcium signaling elicited by glutamate receptor stimulation in developing cerebellar neurons. *Brain Res* 728 (1996): 166-74.
[43] Purohit, V., *et al.* Alcohol, Intestinal Bacterial Growth, Intestinal Permeability to Endotoxin, and Medical Consequences: Summary of a Symposium. *Alcohol* 42 (2008): 349-61.
[44] Carnicella, S., *et al.* GDNF is a fast-acting potent inhibitor of alcohol consumption and relapse. *Proc Natl Acad Sci USA* 105 (2008): 8114-8119.
[45] Braga-Neto, U. M., Hashimoto, R., Dougherty, E. R., Nguyen, D. V., and Carroll, R. J., Is Cross-validation Better than Resubstitution for Ranking Genes**,** *Bioinformatics*, 20 (2004): 253-258.
[46] Agresti, A. *Categorical Data Analysis,* 2nd Edition, New York, Wiley, (2002).

Figures and Tables

Sex
Age
Race
Sexually abused as a child
Otherwise physically abused as a child
Neglected as a child
Experienced sexual trauma
Otherwise experienced physical trauma
Experienced non-physical trauma
Weight
Frequency with which attends religious services
Income
Location of childhood home
Height
Level of education

Table 1: Demographic variables used in the demographic-genetic model and some of the auxiliary models.

1p21.1	1p32.3	3p25.1	4p12	4p21-4p23	5q34	6q25.1
7q31.32	7q33	7q35	10q24.1	11p15.5	11q14.3	11q23.2
11q24.1	12q24.12	13q21.32	15q12	15q25.1	17q24.2	19q13.32

Table 2: Chromosomal regions previously identified as linked to alcoholism.

KEGG Pathways	**p-value**
Calcium Signaling Pathway	0.001
Focal Adhesion	0.002
ECM Receptor Interaction	0.007
Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC)	0.012
Hypertrophic Cardiomyopathy	0.012
Dilated Cardiomyopathy	0.012
Regulation of Actin Cytoskeleton	0.014
Oocyte Meiosis	0.014
Fc-Gamma Receptor-Mediated Phagocytosis	0.021
Long-term Depression	0.036
Adherens Junction	0.040
MAPK Signaling Pathway	0.040
Endocytosis	0.040
Neuroactive Ligand Receptor Interaction	0.047

Table 3: The 14 significant biological pathways ($p < 0.05$) in the demographic-genetic model. 186 total pathways were considered.

Gene	p-value
Intergenic	0.001
BLNK	0.002
BMPER	0.002
SERINC2	0.003
LGALS2	0.004
CPE	0.006
PDLIM5	0.006
PKNOX2	0.008
VEPH1	0.008
NPAS3	0.009
AMPD3	0.010
CADM3	0.010
DAB1	0.010
GLT25D2	0.010

Table 4: The 14 most significant genes in the demographic-genetic model, including the intergenic set. 221 total genes were considered.

Demographic Feature	p-value
Race	0.001
Sex	0.001
Education Level	0.002
Income	0.002

Table 5: The four significant demographic features ($p < 0.05$) in the demographic-genetic model. 15 total demographic features were considered.

Feature-Feature Relationship	p-value
Race-Income	0.011
Sex-rs5933820	0.016
Race-rs8225	0.040
Income-Education Level	0.041

Table 6: The four significant feature-feature relationships ($p < 0.05$) in the demographic-genetic model. 427 total feature-feature relationships were considered.

Controls:

Race	2 C (wild type)	Heterozygous	2 T (variant)
African American	9.2%	41.8%	49.0%
European American	76.3%	22.2%	1.6%

Cases:

Race	2 C (wild type)	Heterozygous	2 T (variant)
African American	6.0%	37.5	56.5%
European American	70.4%	26.5	3.1%

Table 7: Distribution of rs8225 in the two race groups among cases and controls. The link between this variant and “race” group is determined to be significant for predicting alcoholism (see Table 5).

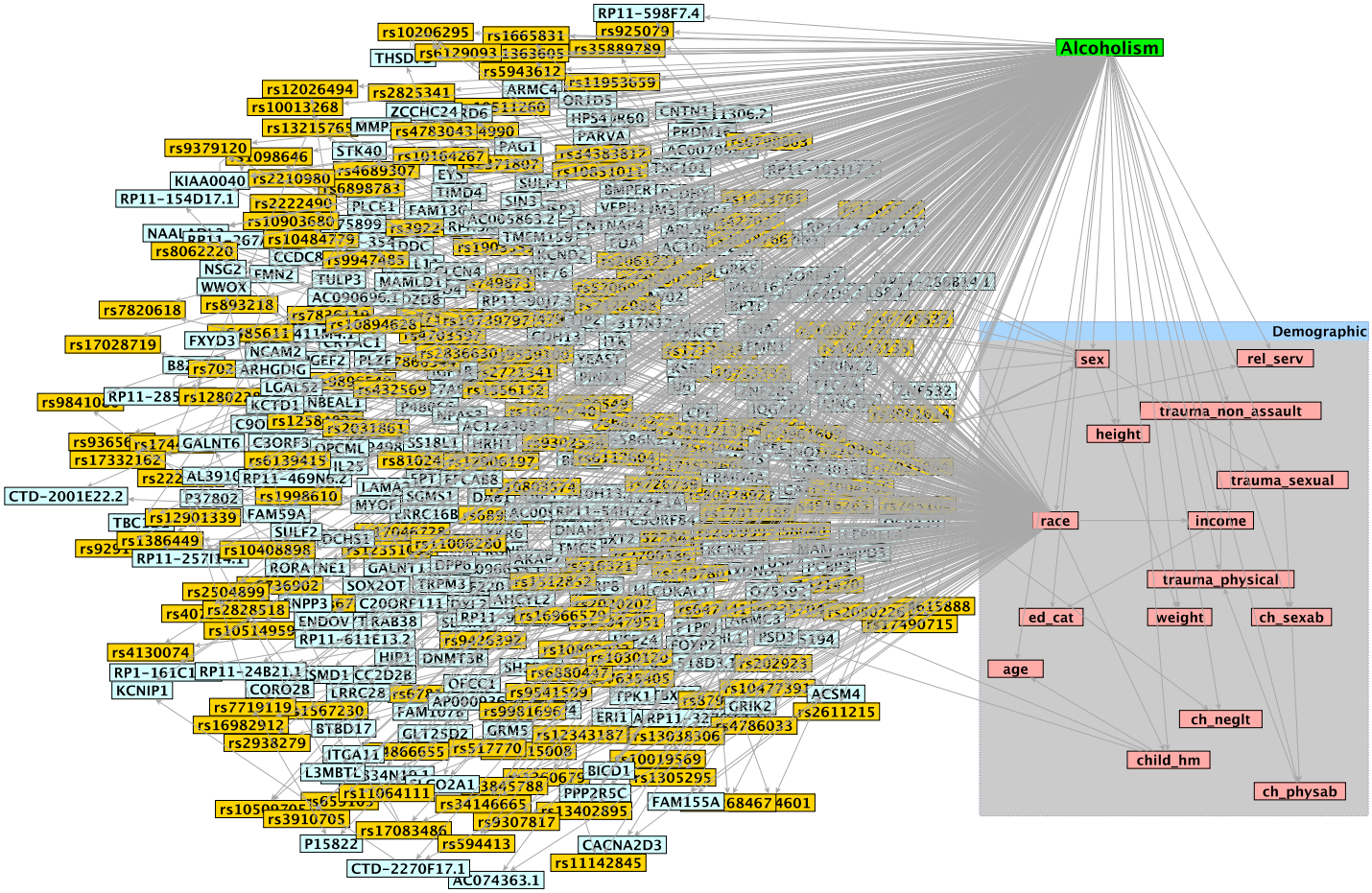

Figure 1: The demographic-genetic prognostic model of Alcoholism. The network contains 413 SNPs and 15 demographic variables. Blue, yellow, and pink boxes represent contributing genes, intergenic SNPs, and demographic variables, respectively. The nodes represent variables and a link between two nodes represents their dependency.

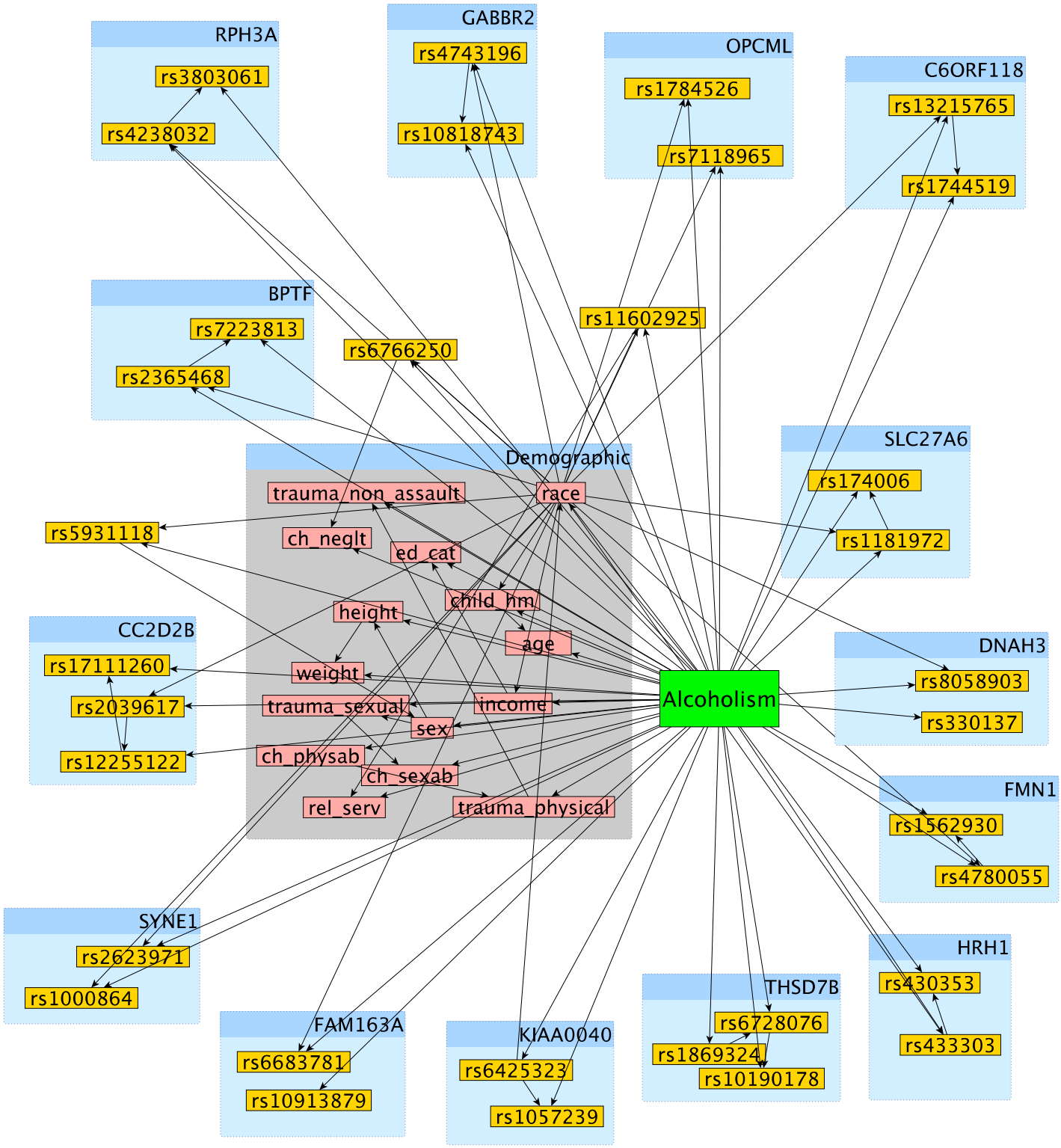

Figure 2: A subgraph of the demographic-genetic Bayesian network. All demographic features are included, as well as the SNPs of several genes that have multiple SNPs in the network. Each blue box is labeled with the gene on which all of the SNPs within the box are found. The pink box contains all of the demographic features.