Predictive Probabilistic Analysis Reveals Multiscale Epistasis in Nicotine Dependence

Amin Zollanvari[123], Aaron Merlob[123], Kent Huynh[4], Jonathan M Dreyfuss[23], Rachel B Ramoni[15], Michael J McGeachie[248], Nancy L Saccone[6], Dorothy K Hatsukami[7], Marco F Ramoni[234], Laura J Bierut[6], Gil Alterovitz[1234*]

[1]Center of Biomedical Informatics, Harvard Medical School, Boston, MA
[2]Children's Hospital Informatics Program at Harvard-MIT Division of Health Science, Boston, MA
[3]Partners Healthcare Center for Personalized Genetic Medicine, Boston, MA
[4]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA
[5]Department of Developmental Biology, Harvard School of Dental Medicine, Boston, MA
[6]Department of Genetics, Washington University School of Medicine, St. Louis, MO
[7]Transdisciplinary Tobacco Use Research Center, Department of Psychiatry, University of Minnesota, Minneapolis, MN
[8]Channing Laboratory, Brigham and Women's Hospital, Boston, MA

*To whom correspondence should be addressed:

gil_alterovitz@hms.harvard.edu.

The search for genetic variants associated with disease traits has epitomized sequence-based studies for nearly a decade. The limited success of genome-wide association studies (GWAS), possibly precipitated by the polygenic nature of the complex traits under study, however, has demonstrated the need for novel, multivariate models capable of quantitatively capturing *interactions* between a host of genetic and non-genetic factors. Here, we present a prediction-based methodology founded on a robust multi-level Bayesian framework to: 1) Capture the multifactor effects of numerous genetic variants and other factors that underlies a complex behavior (i.e. nicotine dependence), 2) Construct a prognostic model for the trait, 3) Delineate the pattern of dependencies among variants, and 4) Identify the pathways, biological processes, single nucleotide polymorphism (SNP) superclasses, and miRNAs with the most significant *predictive* power relative to this behavior in nicotine-exposed individuals. Through use of abstraction and prediction, his methodology is generalizable, and we anticipate seeing it applied to the study of other complex behaviors.

Complex traits are conjectured to originate through frequent, iterative interactions between a large number of environmental and genetic factors, with each interaction contributing relatively little to the final outcome[1]. This conjecture stems from the relatively weak natural selection susceptibility alleles experience[2]. In this scenario, constructing a comprehensive picture of the environmental-genetic architecture underpinning a trait requires a holistic approach that accounts for the effect of many incompletely penetrant variants and environmental factors, and their interactions. Using realistic assumptions about sample sizes, standard statistical tools such as multivariate logistic regression models will be generally incapable of estimating epistasis, i.e. non-additive gene-gene interaction, since its complexity grows exponentially as higher order interactions are added[1,3,4]. For that reason, utilizing holistic methodologies that truly capture the effects of large numbers of variants seems necessary in

gathering a comprehensive picture of the physiopathological processes underpinning a complex trait. Bayesian networks have the potential to efficiently manage this type of complexity[4-6]. We used power of Bayesian networks to not only predict the risk of a complex behavior, but to also identify *predictive* pathways and molecular processes contributing to the trait—the latter *should be* the primary goal of many genome-scale studies[7].

In this work we first used a Bayesian approach to predict the risk of a complex behavior based on Single Nucleotide Polymorphism (SNP) profiles. Then, we used a new predictive-based framework to dissect the constructed network of SNPs/demographics to set of SNPs/demographics that *significantly predict* the phenotype. Through mapping the identified set of SNPs to biological pathways and processes, we obtained a list of pathways/processes that significantly account for predicting the phenotype. In this way we captured the effect that genetic variants, molecular processes, and pathways have on the estimated risk of the complex behavior. Furthermore, we employed this framework to quantify the significance of predictive power of various types of SNPs (intronic, intergenic, synonymous, nonsynonymous, upstream, downstream, 3' UTR, 5' UTR, stop gained, NMD transcript, and within non-coding genes) as well as miRNA-related set of SNPs. We will describe our predictive model of nicotine dependence in Results section A, and the new predictive-based framework for identifying predictive set of SNPs in Results sections C, D, and Discussion sections C and D.

While the proposed Bayesian multi-level analysis is applicable to essentially all GWA studies and sequence-based studies, our primary interest was a complex and common behavior, nicotine dependence. Tobacco use, amplified by nicotine dependence, is responsible for five million deaths annually, making it the leading cause of preventable death worldwide. The death toll from tobacco use will reach 10 million a year by 2020[8]. Nicotine dependence has a strong genetic component[9]; twin studies have demonstrated the heritability of a large proportion of phenotypic variance ranging from 40 – 75%[10] (for more information on nicotine dependence see Supplementary Note 1).

RESULTS

A. A Bayesian Network to Predict Complex Behavior: Nicotine Dependence Application

In order to predict the risk of a complex behavior like nicotine dependence, we applied Bayesian framework for multivariate statistical analysis of human genetic data (See Supplementary Note 2 for more information and a simple example). We employed a tree-augmented structure[6], a subtype of Bayesian network, in order to predict the risk of nicotine dependence for each subject in a multiracial population of European-Americans (EA) and African-Americans (AA), and again in a population of only European-Americans. Figure 1 represents the model built from the multiracial sample, which will henceforth be referred to as the AA+EA model. The AA+EA model is comprised of 426 SNPs and two demographic variables, specifically gender and ethnic group. The SNPs can be further dissected into 345

SNPs located within 5kb of 248 genes and 81 intergenic SNPs for which there was no known gene less than 5kb away. For a compact representation of the results, we clustered SNPs in the same gene into hyper-nodes in Figure 1. The complete list of child-parent sets is given in Supplementary Table 1.
Supplementary Figure 1 shows the Bayesian network constructed on the EA data set (henceforth called the EA model), in which 655 SNPs and gender modulate the risk of nicotine dependence. This network comprises 529 SNPs inside or close to 373 genes and 126 intergenic SNPs. Supplementary Table 2 shows the complete list of child-parent sets in this network. In the present network, all nodes (SNPs and demographic variables) are immediate children of phenotype, yet some nodes have a second parent indicating that a change in the value of the second parent changes the effect the child node has on the phenotype.

B. Characterizing SNPs

Our analysis began by ranking SNPs according to their associations with nicotine dependence. Supplementary Table 1 and 2 show for each SNP in the AA+EA and EA networks, respectively, the numeric rank, the p-value for association, SNP type, the chromosomal position, and the gene or the closet gene.

Figure 2 shows a short region in chromosome 15q24-25.1 (from 78730313bp-78934551bp) where 55 out of 426 SNPs selected for the AA+EA model are located. Since there were only 135 SNPs genotyped in this region, this region was significantly overrepresented ($p<8.69\times10^{-5}$). The genes tagged by these 55 SNPs, are IREB2, AGPHD1, PSMA4, CHRNA3, CHRNB4, and CHRNA5. Several previous studies have confirmed the association between SNPs in the CHRNA5-CHRNA3-CHRNB4 cluster with nicotine dependence[11-13]. That said, Supplementary Tables 3 and 4 demonstrate the limited predictive power of models constructed by considering only SNPs from a single part of the genome. It is particularly relevant to note that chromosome 15q24-25.1, the focus area of many prior studies[11-13], in isolation, delivers an incomplete picture of nicotine dependence in both populations. For more information on the other variables in our model, including race, sex, SNPs, and genes see Supplementary Note 3, 4, and 5. For a visual representation of the implicated genes and interactions in this region see Figure 3 and Supplementary Figure 2.

C. Predictive analysis of molecular processes and pathways

This work utilizes a new framework we developed to identify the effect of biological pathways and molecular processes through a predictive setting (see method section D. for details). Briefly put, in our framework results are generated in two steps: 1) mapping SNPs (or, more generally, arbitrary components of the predictive model) to biological pathways/processes and 2) identifying pathways/processes that statistically significantly predict the risk of nicotine dependence in a population independent from the one in which the predictive-model was constructed. Identified pathways/processes are interpreted as those that either cause or confer protection to the development nicotine dependence, or are altered by mechanisms that cause or confer protection to the development nicotine dependence. For a general

representation of the idea see Figure 4 and Methods section D. The list of biological processes and pathways that significantly predicted nicotine dependence is made available for review in Table 1.

Many of the pathways/processes in Table 1 have been published in various studies and/or for animal models. In this work, we unified all these biological processes and pathways in a single data-driven model. Table 1 lists external evidence that lends support to implication of these biological pathways/processes in nicotine dependence. However, while most are derived from animal models and involve black boxes with nicotine input on one side and a behavioral or biological change quantified, this is the first work to associate many these biological pathways/processes with human nicotine dependence at the genetic level. For example, the effect of nicotine on the *phosphorylation* of an 80-kDa protein in the rat frontal cortex is shown to be related to neuroadaptation and tolerance to nicotine[14]. In another rat model, it was shown that nicotine reverses stabilized *long-term potentiation* in hippocampal CA1 region, which directly affects cognitive functions[15].

Furthermore, using rat models it was demonstrated that not only brain development in adolescence is vulnerable to nicotine[16], but also prenatal nicotine exposure induces brain damage and alters synaptic functionality[17]. An important pathway that effects addiction through synaptic functionality and plasticity is *MAPK signaling,* which can result in long-lasting behavior[18]. Increases of *MAPK* activity in the rat prefrontal cortex are known to be associated with nicotine administration[19]. Furthermore, the *MAPK* pathway plays a crucial role in signaling from nicotinic receptors through its effect on Chromogranin A in rat and mouse[20].

Another notable pathway in Table 1 is the *ErbB signaling.* It has been shown that cigarette smoke activates *NRG1β/ErbB3 signaling,* which in turn activates a variety of signal cascades including *MAPK Signaling pathway*[7]. Nevertheless, *ErbB signaling* pathway is often associated with schizophrenia[21,22]. However, it has reported that patient with schizophrenia have the highest rate of nicotine abuse[23,24]. The significant relationship of *ErbB signaling* pathway with nicotine dependence in our result suggests that this pathway is a candidate for accounting for this high rate of smoking in patients with schizophrenia.

Additionally, nicotine increases the level of nitric oxide in the mouse macrophage cells, which in the presence of *interferon-gamma* (see Table 1) induces relaxation[25]. Interferon-gamma/lipopolysaccharide activation of macrophages enhances cytotoxic effects during inflammation. Conversely, interferon-gamma mediated signaling pathways are a subtype of cytokine-mediated signaling pathway.

Some of the pathways in Table 1 are associated with smoking/nicotine in human studies. For example, it is understood that the smoking behavior depends on sensory stimuli, suggesting a potential route to substitutes for smoking[26]. Sensory cues accompanying inhalation significantly modulate smoker's satisfaction and the reinforcing effect of smoking[26,27]. This can explain the relationship of *olfactory transduction* to nicotine dependence as identified by our model. Some of the effects of smoking and nicotine in human studies are described in Table 1 including: 1) The detrimental effect of smoking on

platelet survival through *blood coagulation*[28], 2) The association of protein *N-linked glycosylation via asparagine* with reinforcing effect of nicotine[29], and 3) The inhibitory effect of nicotine on production of several *cytokines* such as IL-6, IL-12, and TNF-α[30], which has in turn a detrimental effect on human immune system.

These examples demonstrate the ability of our approach to identify mechanisms that affect and/or are affected by the development of nicotine dependence. That said, we cannot directly distinguish between causal relationships and relationships where an outside variable influence both the biological pathways/processes and nicotine dependence.

Figure 5 shows the biological processes and pathways in Table 1. The interrelationships between biological pathways/processes in this figure represent either an interaction pattern between SNPs in the full model or that SNPs that are shared between the two pathways/processes significantly predict the phenotype. The same analysis was used to identify the importance of the predictive power of various types of SNPs (See Supplementary Note 6).

D. Identifying miRNAs

MicroRNAs (miRNAs) are a type of noncoding, single-stranded RNA (typically 22-25 nucleotides in size) with important post-translational regulatory function and implications for developing psychiatric disorders. MicroRNAs are thought to act by affecting synaptic placidity in the brain, which in turn affects cognitive functions. This mechanism of action both aligns well with the neurotropic signaling pathway that we identified in Table 1 and highlights the effect of nicotine on learning, attention, and memory[31]. We again leveraged the prediction-based Bayesian framework used in section C to identify miRNAs that are significantly involved in mechanisms that either a) affect the development of nicotine dependence or b) are altered by nicotine dependence. Supplementary Table 5 provides a list of miRNAs identified under the Bayesian framework. In fact, several miRNAs in this list have, in rat or mouse models, been previously identified as playing a major role in brain and neuron function. In mouse hippocampi, two miRNAs, miR-214 and miR-218, are upregulated by induction of long-term potentiation and depression, respectively[32]. Similarly, in rat hippocampi chronically treated with a mood stabilizer, let-7b and let-7c are downregulated[33]. The miRNAs let-7b and let-7d have been observed to be upregulated in sleep-deprived rat brains[34]. The miRNAs miR-224 and miR-17-3P have shown upregulation as high as 2.18 to 3.20-fold in human gastric adenocarcinoma cells treated with nicotine[35].

E. Validation of the full Models (AA+EA and EA)

We used cross-validation technique to carry out model selection (see ref[4]; also see Methods). In order to validate the AA+EA model, we predicted the occurrence of nicotine dependence in a population of 213 independent individuals composed of 117 cases, 96 controls and 157 European-Americans, 56 African-Americans. The AUROC of the model was 77.4% with 157 out of 213 (73.7%) samples classified

correctly. We similarly validated the EA model, by predicting the occurrence of nicotine dependence in a population of 205 independent European-Americans composed of 106 cases, 99 controls. This model achieved AUROC of 82.3% with 151 out of 205 (73.6%) samples classified correctly.

DISCUSSION

Personalized genetic medicine invokes the promise of informed treatment options for complex diseases or behaviors based on genetic testing, and has been a goal since the time of the first genome-wide association studies[36]. Due to the polygenic nature of complex traits, for each trait, a plethora of loci have been identified that possess strong evidence in favor of association, yet are poor individual predictors. Supplementary Tables 3 and 4 compare accuracies of Bayesian models in terms of prediction via individual SNPs for the fitted data (1), using the cross-validation technique (2), and built on the training data, and applied to the test set (3), using the combined and the EA datasets, respectively. All the AUROCs on the fitted data, attained by single gene analysis over both models, were under 61% (on the borderline of a "fail" in terms of prediction rating[37]). This illustrates the point that each individual SNP conveys only a small amount of information about nicotine dependence.

A. Maximum achievable AUROC

It is noteworthy to compare our results with the AUROC of the best possible genetic test. The theoretical maximum AUROC can be estimated using sibling recurrence risk and trait prevalence. The sibling recurrence risk of nicotine dependence is estimated to be 1.7-2.4 (according to different criteria for nicotine dependence definition[38]), and a prevalence of 24% (ref.[39]). With this in mind, the maximum attainable AUROC for a genetic test that explains all the genetic variance of nicotine dependence can be estimated to be 92%-100% (see ref.[40]). Thus, we find that 'excellent' models can theoretically be developed for this complex as measured by objective metrics[7].

B. Dissecting the full networks to various set of SNPs

As a comparison, we constructed predictive models on the SNPs at nicotinic acetylcholine receptors (nAChRAs) in each model. Specifically, we considered the 219 nicotinic receptor SNPs listed in ref.[11,12]. For the combined model, 69 SNPs were in common between this list and the list of SNPs in our model. A model constructed on these 69 SNPs combined with sex and race variables achieved AUROC of 60.1%. For the European Americans, we constructed another model based on 82 SNPs and sex that achieved an AUROC of 61.1%. We additionally constructed two models on the combined set and EA set, respectively, that considered all 219 SNPs and corresponding demographic variables (sex and/or race). These models achieved AUROC of 61.7% accuracy and 62.4% AUROC on combined set, and on the EA set, respectively. From this, we conclude that the SNPs at nAChRAs, in isolation, are insufficient to reliably predict nicotine dependence, and that other factors (genetic and environmental) are involved in the development of this behavior.

Many of the SNPs previously found to be associated with nicotine dependence are on chromosome 15q24-25.1[11-13]. While this region is significantly enriched in both the AA+EA and EA models, the predictive models based only on variants from this region have poor accuracy. The predictive accuracy of the corresponding models for the combined and EA samples are 63.7% and 64.4% (ranked poor by objective metrics[7]), respectively, as measured by the AUROC (Supplementary Tables 3 and 4). This further suggests that nicotine dependence is a complex disease with important interactions between its causal variants- thus requiring the networked model that we created (which had qualitatively higher accuracies).

C. Predictive analysis of various SNP types

The multi-level prediction analysis that we devised allows us to determine not only the effectiveness of different functional types of SNPs, but also the effect of each biological pathway or process by inspecting the contributing SNPs. Supplementary Figure 3 shows the results of this analysis for various SNPs types in both models. These types are the leaves of the two ontological graphs in Supplementary Figure 3. While there is no functional SNP type that is significantly over- or under-represented (as determined by Fisher's exact test), two types of SNP, namely synonymous and intergenic SNPs, have significant predictive power on nicotine dependence susceptibility (as determined by our analysis). The predictive result for intergenic SNPs aligns well with the emerging view that the majority of the functional sequences do not encode proteins[41]. The same procedure was employed to find miRNAs with a significant predictive contribution, and resulted in identifying a handful of known and novel miRNAs.

D. Predictive analysis of pathways and processes

As mentioned before, the new multi-level predictive framework can be used to quantify the effect of biological processes and pathways as they contribute to nicotine dependence. Through our approach, we identified and unified many processes and pathways that affect nicotine dependence and/or are affected by nicotine. Many of these processes have been previously reported and verified using rat and mouse models; we identified them through our automated data-driven model on human subjects. We believe the kind of analysis carried out here can be a template for other researchers interested in connecting genotype information to the knowledge encoded in biomedical ontologies. This connection not only facilitates the quantification of the effect of more abstract terms presented by ontologies, but also provides a multi-level view of the biological and cellular processes modulating the trait, from the global to the microscopic.

Sequence-based data has been widely touted for identifying new common genetic variants and quantitative risk factors for diseases. There are many publications assessing the effect of a specific region or variant on a complex trait. However, beyond association, these data sources can be used in new prognostic models for complex traits.

E. Advantages of predictive models

Predictive models based on GWA data and other sequence-based data have several advantages in analyzing complex behaviors. First, accurate predictive models assessing the risk of developing a specific phenotype can be constructed directly by combining demographic/environmental and genetic factors. These models are readily translated into clinical practice, e.g. in nicotine dependence application they can be used through counseling and intervention to prevent or stop smoking in individuals with high susceptibility to dependency[36]. Next, predictive measures generally do not have common analysis issues such as multiple testing or multicollinearity that prevent genetic association tests from being replicated or applicable to future samples[3]. This means that if a genetic/environmental model has good predictive power (inferred from estimates of the true predictive power), the model is generalizable to future samples. Lastly, comprehensive predictive modeling can capture the complex interactions between many demographic/environmental and genetic factors, providing a characterization of the underlying biological and environmental mechanism that determines the phenotype.

METHODS

A. Data Collection

As a part of the Collaborative Genetic Study of Nicotine Dependence (COGEND), individuals were recruited through a telephone screening of subjects to determine eligibility for recruitment with either case or control status. The individuals have reported lifetime smoking of at least 100 cigarettes, which is the threshold traditionally used to define a smoker[42]. Cases are considered nicotine dependent according to the Fagerström Test for Nicotine Dependence (FTND)[43] with an FTND score of 4 or more. Controls were never nicotine dependent and had an FTND score of 0 or 1. The 2772 study participants comprise 1524 cases and 1248 controls across two ethnic groups where 2062 are European-American and 710 are African-American. These subjects were assayed at 1642 SNPs from genes with prior knowledge in favor of their contribution to nicotine dependence. Following another COGEND study[11], the samples of African-American and European Americans were pooled together and the association analysis were performed on the combined sample as explained in the following section. We described further details on data collection elsewhere[11,12].

B. Statistical analysis

Here, the statistical analysis of the AA+EA model is described. The same analysis was performed for the EA model by replacing the 2772 combined sample data set with 2062 European Americans and replacing the Cochran-Mantel-Haenszel statistic with the Cochran-Armitage test statistic[44] for the association analysis.

To avoid any possible artifacts induced by imputation, we removed all those SNPs with a genotyping rate of less than 90%. After this filtering, the average genotype rate on the 1573 remaining SNPs was 99.6%.

After this initial quality control, we randomly split the data set into two parts with 2559 samples as training data and 213 samples set aside for validating our model. The training data was used in association analysis, feature selection and training the Bayesian network and the 213 validation samples were only used at the end of all these analyses for assessing the performance of the model (for the EA model 205 samples were set aside). After this split, there were 2 SNPs in the training data with a genotyping rate of 90%-95% and the rest were over 95%. After imputation on the training samples (which contained representatives from both ethnic groups), we employed the Cochran-Mantel-Haenszel test statistic to rank the SNPs according to their association to nicotine dependence while controlling for ethnicity (test implemented in PLINK v1.06[45]). We chose an initial list of SNPs for use in constructing the best Bayesian network with p<0.1, resulting in 706 SNPs (the top 526 SNPs had p<0.05). This initial set of features with p<0.1 had 807 SNPs for EA model (the top 665 SNPs had p<0.05). Then we constructed a Tree-Augmented Naive Bayesian (TAN) network[6], an extension to the maximal weighted spanning tree algorithm proposed by Chow and Liu[46]. Evidence shows that in a high-dimensional space this classifier generally outperforms many commonly used classifiers, including naïve Bayes networks[6,47], which is incapable of inferring the pattern of dependencies between variants. TAN networks lend themselves to capturing weak dependencies among the features given the classification node. TAN networks can capture dependencies between features, like a more complicated general Bayesian network, but are more efficient in discovering the optimal structure. All together, we believe TAN is a potential candidate for analyzing many complex traits like nicotine dependence.

The maximal weighted spanning tree algorithm, which is the core algorithm to construct the TAN structure, is guaranteed to find the tree structure that maximizes the likelihood of data when the underlying distribution of the data is a tree dependence structure (meaning each node has a single parent and one node (the root) has no parent). In this scenario, the likelihood of data, $D = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}\}$, in which $\mathbf{x_i}$, $i = 1,2,\dots,N$, are independent samples of n features denoted by $\mathbf{x_i} = (\mathbf{x_i^1}, \mathbf{x_i^2}, \dots, \mathbf{x_i^n})$, can be written as:

$$L(D) = \prod_{i=1}^{N} P(\mathbf{x_i}) = \prod_{i=1}^{N} P\left(\mathbf{x_i^{k_j}} | \mathbf{x_i^{k_{m_j}}}\right) \quad (1)$$

in which the tuple $(\boldsymbol{k_1}, \boldsymbol{k_2}, \dots, \boldsymbol{k_n})$ is an unknown permutation of integers $1,2,\dots,n$; m_j an integer such that $0 \leq m_j < j$ (showing the parent index for j), $P(\mathbf{x_j})$ is just probability of $\mathbf{x_j}$, and $P(\mathbf{x_j}|\mathbf{x_0}) = 0$.

Equation (1) is just the mathematical representation of a tree-type dependency structure in which each node can have only one parent and the final "best" estimated structure does not depend on any ordering of the features. If we define the best structure to be the one that maximizes $L(D)$ in (1) over all possible tree structures, called *T*, then after some mathematical manipulations, it can be shown that[46]:

$$\max_T L(D) = \max_T \sum_{j=1}^{n} \hat{I}(\mathbf{x}_i^{k_j}, \mathbf{x}_i^{k_{m_j}}) + C \quad (2)$$

in which C is independent of the choice of the structure and $\hat{I}(\mathbf{x}_i^{k_j}, \mathbf{x}_i^{k_{m_j}})$ is just the estimated mutual information between $\boldsymbol{k_j}$ and its parent $\boldsymbol{k_{m_j}}$. Mutual information between two variables *x* and *y* is defined to be $I(x, y) = \sum_{x,y} P(x, y) \log \frac{p(x,y)}{p(x)p(y)}$ and is the amount of information *y* carries about *x*. Therefore, it can be seen from (2) that to maximize the likelihood function defined over the space of all possible trees, it suffices to maximize the sum of the mutual information, and, in turn, maximize the mutual information between every node. This procedure was extended to maximize this summation in order to find the tree augmented structure in which every node has the class label (phenotype) as one parent and the node with maximum mutual information as the other parent, conditional on the tree-structure[6].

C. Classifier Design and Predictive Validation

We used the 706 previously selected SNPs as the initial list of SNPs to design the TAN (807 SNPs for EA model). A model was then designed and its performance was assessed using cross-validation. Next, a smaller subset of features was selected by removing the 5 least-important features from the bottom of the list of associations. Then, a new model was designed and performance reassessed. This procedure was repeated until few features remained, causing poor performance. All models contained the ethnicity and gender nodes. The best model identified in this way was the model with the best-estimated accuracy on training data using cross-validation. The best model identified in this way was then validated on the set of independent individuals that were randomly set aside from the original data. The AUROC we obtained was 77.4% (82.3% for the EA model).

D. Identifying ontological concepts, different type of SNPs, and miRNAs with significant predictive power

We mapped the genes in our constructed network (i.e. AA+EA network) to the most specific terms in the biological processes branch of GO and also to the pathways in KEGG. In this way we obtained the set of SNPs corresponding to each ontological concept, which in turn represents a biological process or pathway. Then, using each SNP set, a Bayesian network was constructed to predict the risk of development of the behavioral phenotype (here, nicotine dependence). We recorded the predictive power of each SNP set (for each concept) by evaluating its predictive power as measured by the AUROC on the set of independent individuals, described in Methods section B. Then, for each set, we randomly selected the same number of SNPs from the set of 426 SNPs combined with gender and ethnicity and recorded the AUROC of the corresponding Bayesian network. The process of randomly selecting SNPs was repeated 1,000 times for each SNP set. Using 1,000 AUROCs for each set, we estimated the p-value of the observed predictive model of the SNP set. For GO, we had 191 concepts (hypotheses) with the associated number of SNPs varying between 2 to 74. For KEGG we had 51 pathways corresponding to our genes. Using the FDR correction for

multi-hypothesis tests (FDR<0.05) we identified only 4 KEGG pathways. However, as neither single variants nor a small number of variants can account for phenotypic variation of complex traits (as they have small to moderate effect size), it is natural to believe that the same argument holds for pathway and/or processes. Therefore, in the results reported in Table 1 and 2 we used a more liberal significance level of 0.05 without correction for multi-hypothesis tests. The same procedure was performed to see the effect of different type of SNPs (intronic, intergenic, synonymous, nonsynonymous, upstream, downstream, 3' UTR, 5' UTR, stop gained, NMD transcript, and within non-coding genes) as well as miRNAs (we had 127 hypotheses to test that each corresponds to a gene-set that share a miRNA's binding motifs). The mappings between genes and GO concepts, KEGG, SNP type, and miRNAs are available from databases in GO[48], MSigDB[49], Ensembl[50], and MSigDB[49], respectively.

SUPPLEMENTARY MATERIAL

Supplementary Notes file provides additional information regarding nicotine dependence, Bayesian networks, and the AA+EA and EA models.

Supplementary Figure 2 is the counterpart of Figure 3a for the EA model.

Supplementary Figure 3 is the result of traditional enrichment analysis for different functional type of SNPs.

Supplementary Figure 4 is the result of traditional enrichment analysis for different functional type of SNPs.

Supplementary Table 1 and 2 shows each SNP and its parent, position, the p-values for its association, functionality, and the gene or the closest gene for each node in the AA+EA and EA model, respectively.

Supplementary Table 3 and 4 show the accuracy of the models designed in different scenarios for the combined and EA samples, respectively.

Supplementary Table 5 shows the identified mi-RNAs using predictive-based analysis of miRNA-related SNP sets.

ACKNOWLEDGMENTS

In memory of Theodore Reich, founding Principal Investigator of COGEND, we are indebted to his leadership in the establishment and nurturing of COGEND and acknowledge with great admiration his seminal scientific contributions to the field. Lead investigators directing data collection are Laura Bierut, Naomi Breslau, Dorothy Hatsukami, and Eric Johnson. The authors thank Heidi Kromrei and Tracey Richmond for their assistance in data collection. This work was supported by the NIH grants P01CA89392 from the National Cancer Institute, 5R21DA025168-02 (G. Alterovitz), 1R01HG004836-01(G. Alterovitz), and 4R00LM009826¬03 (G. Alterovitz).

REFERENCES

1. Zondervan, K.T. & Cardon, L.R. The complex interplay among factors that influence allelic association. *Nature reviews. Genetics* **5**, 89-100 (2004).
2. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics* **6**, 95-108 (2005).
3. Heidema, A.G. *et al.* The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC genetics* **7**, 23 (2006).
4. Sebastiani, P., Ramoni, M.F., Nolan, V., Baldwin, C.T. & Steinberg, M.H. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nature genetics* **37**, 435-40 (2005).
5. Friedman, N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* **303**, 799-805 (2004).
6. Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian Network Classifiers. *Machine Learning* **29**, 131-163 (1997).
7. Hirschhorn, J.N. Genomewide association studies--illuminating biologic pathways. *N Engl J Med* **360**, 1699-701 (2009).
8. World Health Organization. Report on the Global Tobacco Epidemic. (2008).
9. Swan, G.E. *et al.* A genome-wide screen for nicotine dependence susceptibility loci. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **141B**, 354-60 (2006).
10. Broms, U., Silventoinen, K., Madden, P.A., Heath, A.C. & Kaprio, J. Genetic architecture of smoking behavior: a study of Finnish adult twins. *Twin research and human genetics : the official journal of the International Society for Twin Studies* **9**, 64-72 (2006).
11. Saccone, N.L. *et al.* Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. *Genes, brain, and behavior* **9**, 741-50 (2010).
12. Saccone, N.L. *et al.* The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer research* **69**, 6848-56 (2009).
13. Thorgeirsson, T.E. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638-42 (2008).
14. Ochoa, E.L. & O'Shea, S.M. Concomitant protein phosphorylation and endogenous acetylcholine release induced by nicotine: dependency on neuronal nicotinic receptors and desensitization. *Cell Mol Neurobiol* **14**, 315-40 (1994).
15. Guan, X., Nakauchi, S. & Sumikawa, K. Nicotine reverses consolidated long-term potentiation in the hippocampal CA1 region. *Brain Res* **1078**, 80-91 (2006).
16. Trauth, J.A., Seidler, F.J. & Slotkin, T.A. An animal model of adolescent nicotine exposure: effects on gene expression and macromolecular constituents in rat brain regions. *Brain Res* **867**, 29-39 (2000).
17. Lichtensteiger, W., Ribary, U., Schlumpf, M., Odermatt, B. & Widmer, H.R. Prenatal adverse effects of nicotine on the developing brain. *Prog Brain Res* **73**, 137-57 (1988).
18. Gould, T.J. Nicotine and hippocampus-dependent learning: implications for addiction. *Mol Neurobiol* **34**, 93-107 (2006).
19. Konu, O. *et al.* Region-specific transcriptional response to chronic nicotine in rat brain. *Brain Res* **909**, 194-203 (2001).

20. Tang, K., Wu, H., Mahata, S.K. & O'Connor, D.T. A crucial role for the mitogen-activated protein kinase pathway in nicotinic cholinergic signaling to secretory protein transcription in pheochromocytoma cells. *Mol Pharmacol* **54**, 59-69 (1998).
21. Buonanno, A. The neuregulin signaling pathway and schizophrenia: from genes to synapses and neural circuits. *Brain Res Bull* **83**, 122-31 (2010).
22. Neddens, J., Vullhorst, D., Paredes, D. & Buonanno, A. Neuregulin links dopaminergic and glutamatergic neurotransmission to control hippocampal synaptic plasticity. *Commun Integr Biol* **2**, 261-4 (2009).
23. Goff, D.C., Henderson, D.C. & Amico, E. Cigarette smoking in schizophrenia: relationship to psychopathology and medication side effects. *Am J Psychiatry* **149**, 1189-94 (1992).
24. Hughes, J.R., Hatsukami, D.K., Mitchell, J.E. & Dahlgren, L.A. Prevalence of smoking among psychiatric outpatients. *Am J Psychiatry* **143**, 993-7 (1986).
25. Chen, Y.C., Shen, S.C., Lin, H.Y., Tsai, S.H. & Lee, T.J. Nicotine enhancement of lipopolysaccharide/interferon-gamma-induced cytotoxicity with elevating nitric oxide production. *Toxicol Lett* **153**, 191-200 (2004).
26. Bryant, B., Xu, J., Audige, V., Lischka, F.W. & Rawson, N.E. Cellular Basis for the Olfactory Response to Nicotine. *ACS Chemical Neuroscience* **1**, 246-256 (2010).
27. Rose, J.E., Tashkin, D.P., Ertle, A., Zinser, M.C. & Lafer, R. Sensory blockade of smoking satisfaction. *Pharmacol Biochem Behav* **23**, 289-93 (1985).
28. Singh, J.M. & Singh, M.D. Alkaloids of tobacco and blood coagulation: effect of nicotine on thrombin and fibrinogen. *Clin Toxicol* **8**, 43-52 (1975).
29. Zhang, L., Kendler, K.S. & Chen, X. The mu-opioid receptor gene and smoking initiation and nicotine dependence. *Behav Brain Funct* **2**, 28 (2006).
30. Ouyang, Y. *et al.* Suppression of human IL-1beta, IL-2, IFN-gamma, and TNF-alpha production by cigarette smoke extracts. *J Allergy Clin Immunol* **106**, 280-7 (2000).
31. Couey, J.J. *et al.* Distributed network actions by nicotine increase the threshold for spike-timing-dependent plasticity in prefrontal cortex. *Neuron* **54**, 73-87 (2007).
32. Park, C.S. & Tang, S.J. Regulation of microRNA expression by induction of bidirectional synaptic plasticity. *J Mol Neurosci* **38**, 50-6 (2009).
33. Zhou, R. *et al.* Evidence for selective microRNAs and their effectors as common long-term targets for the actions of mood stabilizers. *Neuropsychopharmacology* **34**, 1395-405 (2009).
34. Davis, C.J., Bohnet, S.G., Meyerson, J.M. & Krueger, J.M. Sleep loss changes microRNA levels in the brain: a possible mechanism for state-dependent translational regulation. *Neurosci Lett* **422**, 68-73 (2007).
35. Shin, V.Y. *et al.* NF-kappaB targets miR-16 and miR-21 in gastric cancer: involvement of prostaglandin E receptors. *Carcinogenesis* **32**, 240-5 (2011).
36. Herbert, L.J., Walker, L.R., Sharff, M.E., Abraham, A.A. & Tercyak, K.P. Are adolescents with ADHD interested in genetic testing for nicotine addiction susceptibility? *International journal of environmental research and public health* **7**, 1694-707 (2010).
37. Pines, J.M. & Everett, W.W. *Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules*, (Blackwell, 2008).
38. Niu, T. *et al.* Nicotine dependence and its familial aggregation in Chinese. *Int J Epidemiol* **29**, 248-52 (2000).
39. Breslau, N., Johnson, E.O., Hiripi, E. & Kessler, R. Nicotine dependence in the United States: prevalence, trends, and smoking persistence. *Arch Gen Psychiatry* **58**, 810-6 (2001).
40. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* **6**, e1000864 (2010).
41. Lander, E.S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187-97 (2011).
42. Centers for Disease Control and Prevention. Tobacco use among adults -- United States: 2005. Morbidity and Mortality. Vol. 55 1145-48 (2006).

43. Heatherton, T.F., Kozlowski, L.T., Frecker, R.C. & Fagerstrom, K.O. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *British journal of addiction* **86**, 1119-27 (1991).
44. Agresti, A. *Categorical data analysis*, (Wiley, New York, 2002).
45. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-75 (2007).
46. Chow, C.K. & Liu, C.N. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory* **14**, 462-467 (1968).
47. Cheng, J. & Greiner, R. Comparing Bayesian Network Classifiers. in *Fifteenth international conference on uncertainty in artificial intelligence* 101-108 (Morgan Kaufmann Publishers, 1999).
48. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
49. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
50. Hubbard, T.J. *et al.* Ensembl 2007. *Nucleic Acids Res* **35**, D610-7 (2007).

Figures/Tables Captions:

Figure 1. The prognostic model of nicotine dependence within two ethnic groups: African-Americans and European-Americans (AA+EA model). The model comprises 426 SNPs and two demographic variables, gender and ethnic group. This network dissects the risk of nicotine dependence to 81 intergenic SNPs and 345 SNPs that are clustered in 248 genes from which 36 genes contain at least two. The light blue boxes are the 36 genes with at least two SNPs inside. The interaction pattern among SNPs in these hypernodes, as well as other SNPs in the model, can be found in Supplementary Table 1. The orange boxes are intergenic SNPs with the labels showing the rs IDs. The cyan boxes show the genes with are selected with only one SNP. The labels for these boxes are genes. The two green boxes show the demographic variables, SEX and ETHNICITY. The phenotype is indicated by the yellow box. The orange oval represents chr15q24-25.1 that has been the main focus of research in nicotine dependence. Although this chromosomal region is enriched (Figure 2), but it has limited power to predict nicotine dependence and hence, we have to consider many other regions and dependencies as presented in this figure (see supplementary Table 3 and 4).

Figure 2. Enrichment of the region 15q24-25.1 for its association with nicotine dependence is shown here. On top is the spread of 426 selected SNPs across genome. At the bottom is the region where 55 SNPs are located (787303133bp-78934551bp) indicating this region is highly associated with nicotine dependence ($<8.69\times10^{-5}$). Y axis is $-\log_{10}$ of p-values for CMH test. The genes tagged by these 55 SNPs, are IREB2 iron-sensing response element (10 SNPs), AGPHD1, of unclear functionality (also known as LOC123688) (8 SNPs), PSMA4, Proteasome degradation functioning (2 SNPs), and three nicotinic acetylcholine receptor subunits, CHRNA3 (15 SNPs), CHRNB4 (11 SNPs), CHRNA5 (8 SNPs), and 1 intergenic SNP 6247bp upstream of CHRNA5. Here we have counted SNPs between 5kbp upstream or downstream of a gene.

Figure 3. (a) A sub-network taken from the original network in Figure 1. The figure shows the within-gene and between-gene SNP-SNP interactions in several previously reported genes as being associated with nicotine dependence. Each gene is a "box" with the SNPs shown inside. SNPs that are physically located in an intergenic region close to these genes (<25kbp) are shown by nodes close to the "gene-box". Most identified interactions are across SNPs inside genes, and then across genes on the same chromosome, e.g. CHRNB1 and ZBTB4, both located at 17p13.1, CHRNB4, CHRNA3, CHRNA5, PSMA4, AGPHD1 and IREB2 all located at 15q24-25.1, and CHRNB3 located at 8p11.21. **(b)** The implicated gene-gene interaction on chromosome 15q24-25.1 (consistent across AA+EA model and EA model). As we observe in this figure CHRNA5 is directly interacting with four other genes, suggesting the importance of this gene at chromosome 15q24-25.1.

Figure 4. Multiscale epistasis in nicotine dependence. The network, inferred from SNP profiles, is at the bottom. The network of genes is in the middle. The corresponding biological pathways and molecular processes are on top. The effect of each pathway/process is evaluated by taking all the data-dependent entities (SNPs, genes, proteins,…) corresponding to that and evaluating the significance of AUROC for the corresponding predictive model (Bayesian network here). The mapping between pathways/processes to the corresponding data-dependent entities depends on available ontologies/knowledge bases. Here in order to find such mappings we used the available correspondence of pathways/processes to genes combined with genes to SNPs mappings. Constructing accurate ontologies in the future will improve the performance of our framework in identifying predictive pathways/processes.

Figure 5. Molecular processes and pathways that significantly predict nicotine dependence. We assume two biological pathways, molecular processes, etc. have statistical dependency if they share SNPs in some common genes between them or there is statistical dependency across corresponding SNPs (through their harboring genes) inferred from Bayesian network. However, we only report those statistical dependencies that significantly contribute to prediction of the phenotype. The black links indicate the association of the pathways (represented by yellow nodes) and processes (light blue nodes) with nicotine dependence. The red arrows indicate the interrelationship among pathways and biological processes that significantly differentiate nicotine dependent individuals from non-dependent individuals.

Table 1. List of identified biological pathways/processes, the p-value for predictive power of the biological pathway/process-related set of SNPs, and the number of pathway/process-related set of SNPs in the network of Figure 1. The related set of SNPs for each pathway/process is defined as the SNPs with the distance less than 5kbp upstream or downstream from the genes contributing to each pathway/process. Ontologies and knowledge bases can be used to determine the contributing genes to each pathway/process.

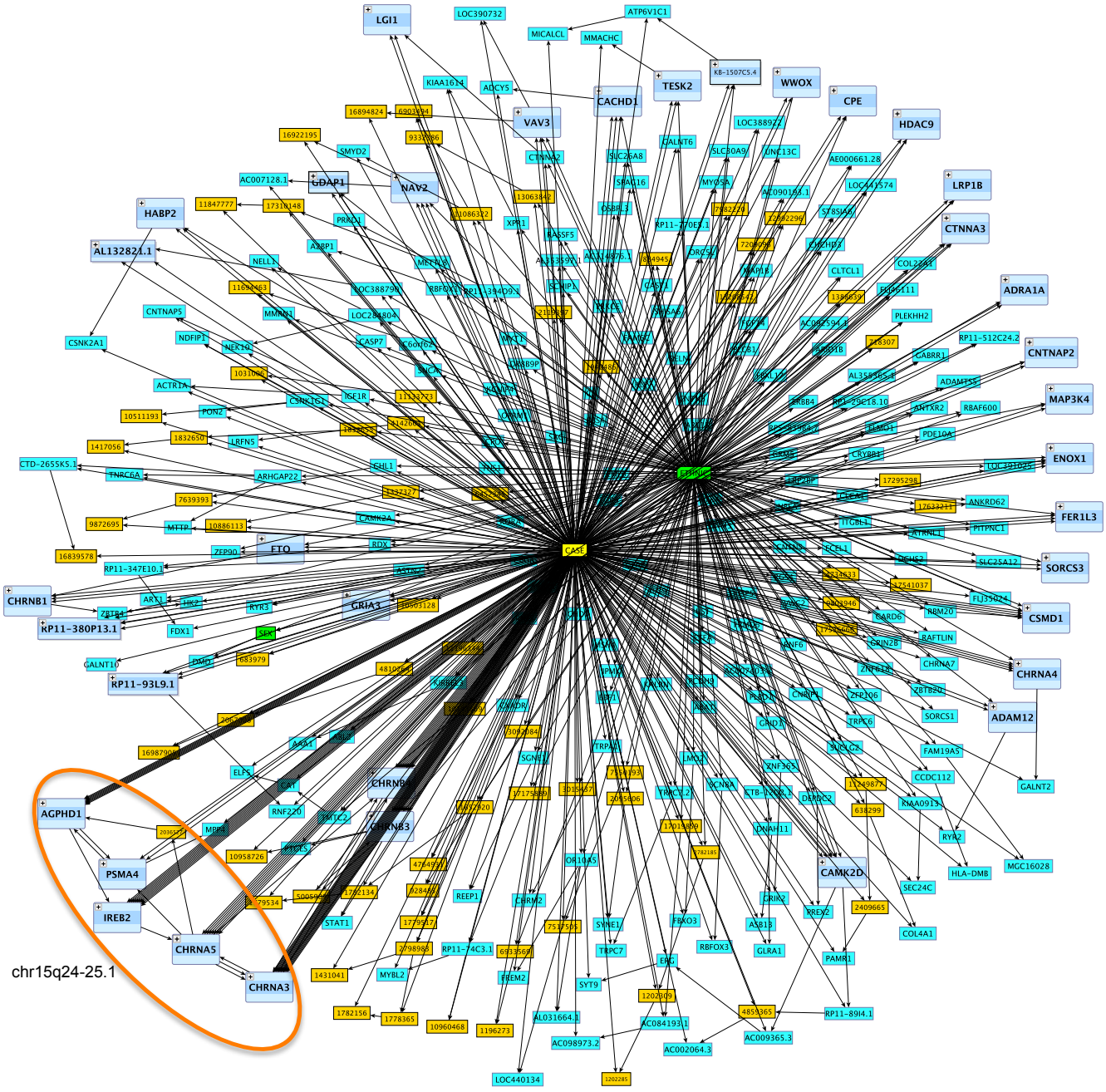

Figure 1. The prognostic model of nicotine dependence within two ethnic groups: African-Americans and European-Americans (AA+EA model). The model comprises 426 SNPs and two demographic variables, gender and ethnic group. This network dissects the risk of nicotine dependence to 81 intergenic SNPs and 345 SNPs that are clustered in 248 genes from which 36 genes contain at least two. The light blue boxes are the 36 genes with at least two SNPs inside. The interaction pattern among SNPs in these hypernodes, as well as other SNPs in the model, can be found in Supplementary Table 1. The orange boxes are intergenic SNPs with the labels showing the rs IDs. The cyan boxes show the genes with are selected with only one SNP. The labels for these boxes are genes. The two green boxes show the demographic variables, SEX and ETHNICITY. The phenotype is indicated by the yellow box. The orange oval represents chr15q24-25.1 that has been the main focus of research in nicotine dependence. Although this chromosomal region is enriched (Figure 2), but it has limited power to predict nicotine dependence and hence, we have to consider many other regions and dependencies as presented in this figure (see supplementary Table 3 and 4).

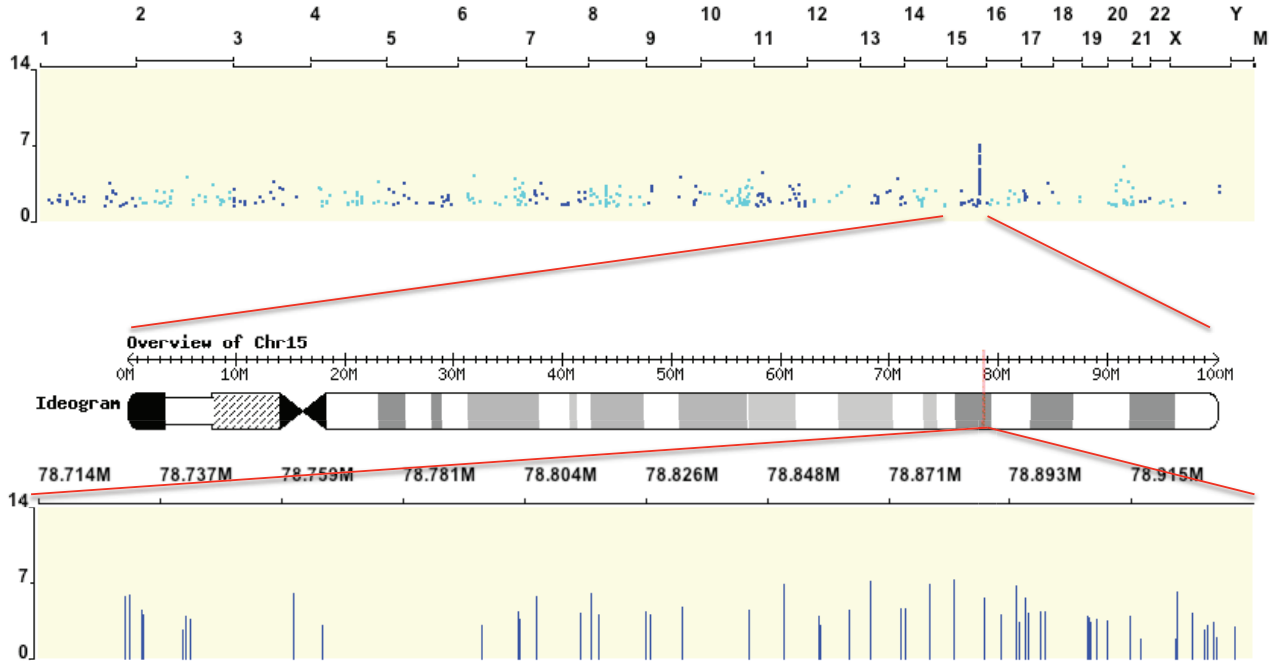

Figure 2. Enrichment of the region 15q24-25.1 for its association with nicotine dependence is shown here. On top is the spread of 426 selected SNPs across genome. At the bottom is the region where 55 SNPs are located (787303133bp-78934551bp) indicating this region is highly associated with nicotine dependence ($<8.69\times10^{-5}$). Y axis is $-\log_{10}$ of p-values for CMH test. The genes tagged by these 55 SNPs, are IREB2 iron-sensing response element (10 SNPs), AGPHD1, of unclear functionality (also known as LOC123688) (8 SNPs), PSMA4, Proteasome degradation functioning (2 SNPs), and three nicotinic acetylcholine receptor subunits, CHRNA3 (15 SNPs), CHRNB4 (11 SNPs), CHRNA5 (8 SNPs), and 1 intergenic SNP 6247bp upstream of CHRNA5. Here we have counted SNPs between 5kbp upstream or downstream of a gene.

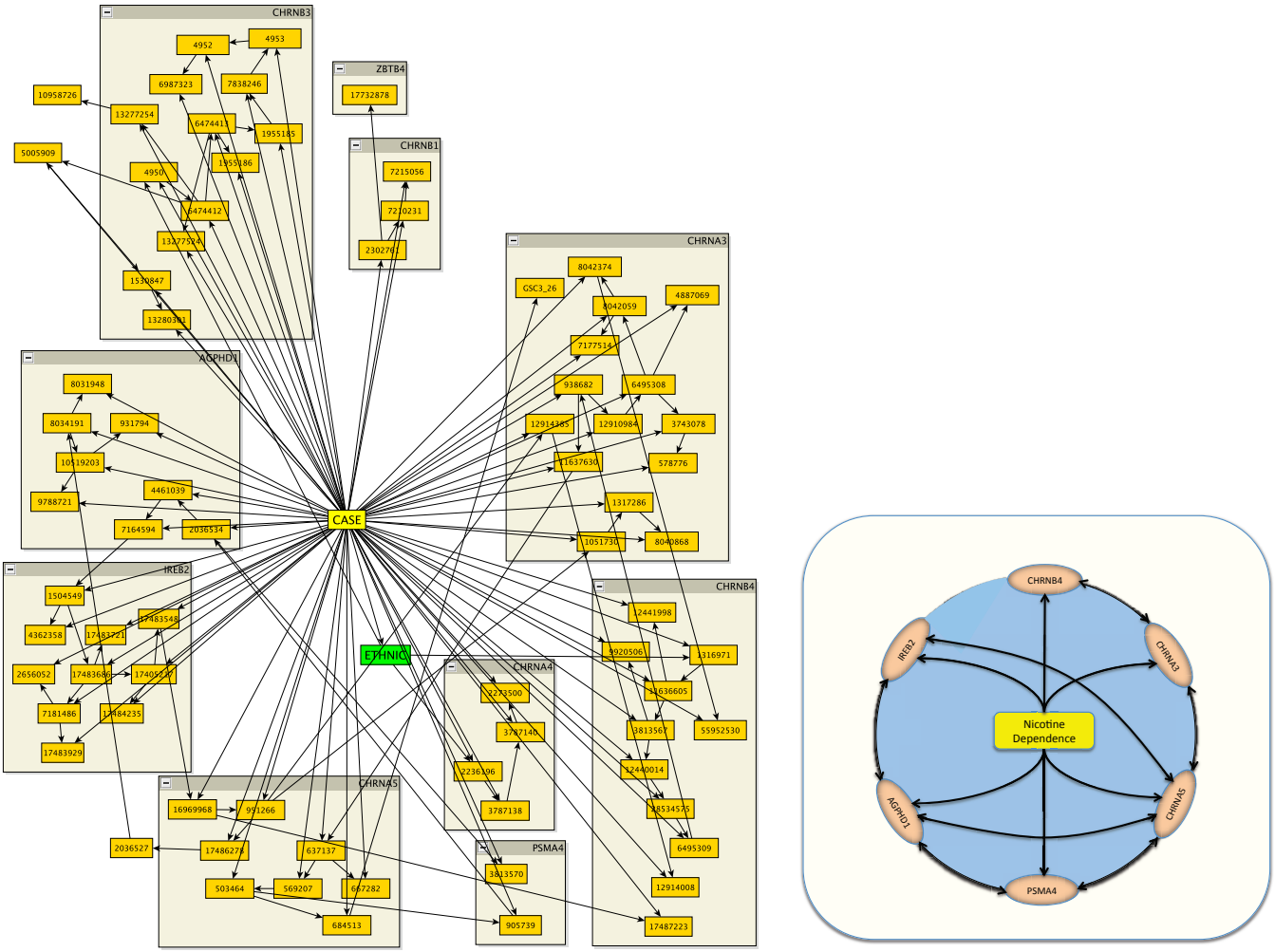

Figure 3. (a) A sub-network taken from the original network in Figure 1. The figure shows the within-gene and between-gene SNP-SNP interactions in several previously reported genes as being associated with nicotine dependence. Each gene is a "box" with the SNPs shown inside. SNPs that are physically located in an intergenic region close to these genes (<25kbp) are shown by nodes close to the "gene-box". Most identified interactions are across SNPs inside genes, and then across genes on the same chromosome, e.g. CHRNB1 and ZBTB4, both located at 17p13.1, CHRNB4, CHRNA3, CHRNA5, PSMA4, AGPHD1 and IREB2 all located at 15q24-25.1, and CHRNB3 located at 8p11.21. **(b)** The implicated gene-gene interaction on chromosome 15q24-25.1 (consistent across AA+EA model and EA model). As we observe in this figure CHRNA5 is directly interacting with four other genes, suggesting the importance of this gene at chromosome 15q24-25.1.

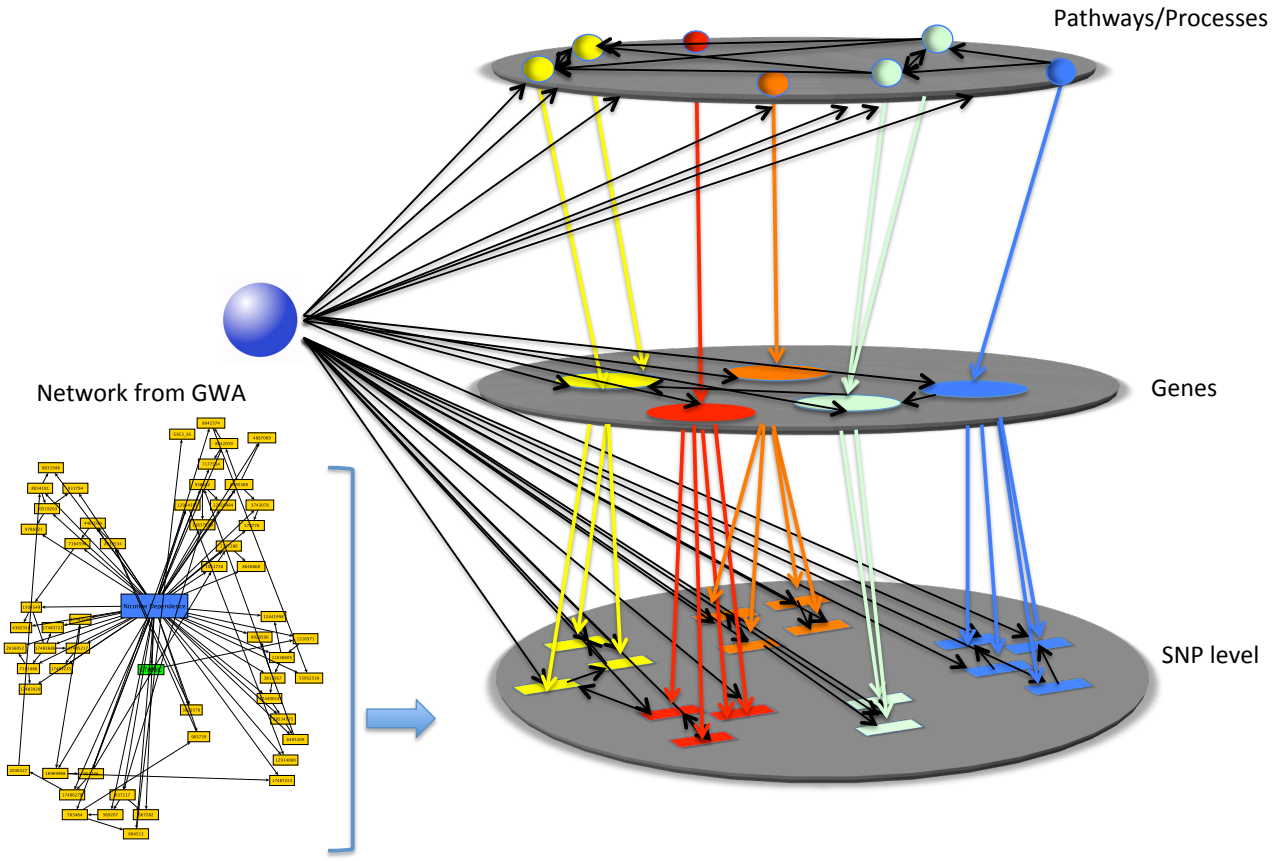

Figure 4. Multiscale epistasis in nicotine dependence. The network, inferred from SNP profiles, is at the bottom. The network of genes is in the middle. The corresponding biological pathways and molecular processes are on top. The effect of each pathway/process is evaluated by taking all the data-dependent entities (SNPs, genes, proteins, etc.) corresponding to that and evaluating the significance of AUROC for the corresponding predictive model (Bayesian network here). The mapping between pathways/processes to the corresponding data-dependent entities depends on available ontologies/knowledge bases. Here in order to find such mappings we used the available correspondence of pathways/processes to genes combined with genes to SNPs mappings. Constructing accurate ontologies in the future will improve the performance of our framework in identifying predictive pathways/processes.

Concept	p-value	#SNPs
Protein Phosphorylation	0.022	10
Inactivation of MAPK Activity	0.002	2
Protein N-linked Glycosylation via Asparagine	0.008	2
Positive Regulation of Rac GTPase Activity	0.043	2
Post-translational Protein Modification	0.008	2
Chloride Transport	0.017	3
Regulation of Cell Growth	0.047	4
Blood Coagulation	0.01	7
Interferon-gamma-mediated Signaling Pathway	0.031	7
Leukocyte Transendothelial Migration	0.001	7
Olfactory Transduction	0.001	7
Viral Myocarditis	0.002	4
Arrhythmogenic Right Ventricular Cardiomyopathy	0.003	5
Tight Junction	0.006	6
Pathways in Cancer	0.016	8
Wnt Signaling Pathway	0.023	7
Long Term Potentiation	0.026	8
FC Gamma R Mediated Phagocytosis	0.028	4
FC Epsilon RI Signaling Pathway	0.032	3
Glioma	0.032	6
Adherens Junction	0.033	5
Neurotrophin Signaling Pathway	0.039	5
ErbB Signaling Pathway	0.04	7
Endometrial Cancer	0.04	3
Systemic Lupus Erythematosus	0.043	3

Table 1. List of identified biological pathways/processes, the p-value for predictive power of the biological pathway/process-related set of SNPs, and the number of pathway/process-related set of SNPs in the network of Figure 1. The related set of SNPs for each pathway/process is defined as the SNPs with the distance less than 5kbp upstream or downstream from the genes contributing to each pathway/process. Ontologies and knowledge bases can be used to determine the contributing genes to each pathway/process.

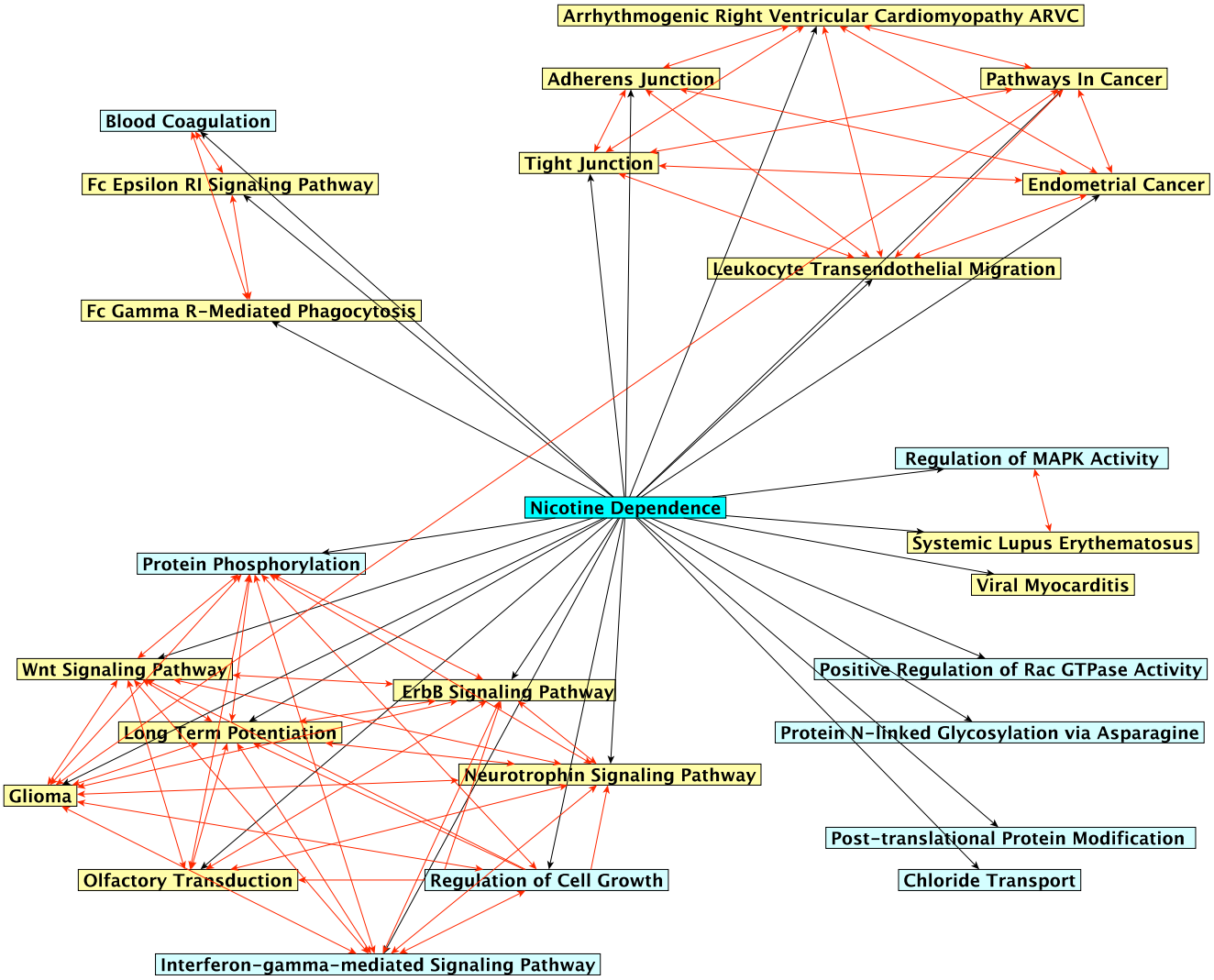

Figure 5. Molecular processes and pathways that significantly predict nicotine dependence. We assume two biological pathways, molecular processes, etc. have statistical dependency if they share SNPs in some common genes between them or there is statistical dependency across corresponding SNPs (through their harboring genes) inferred from Bayesian network. However, we only report those statistical dependencies that significantly contribute to prediction of the phenotype. The black links indicate the association of the pathways (represented by yellow nodes) and processes (light blue nodes) with nicotine dependence. The red arrows indicate the interrelationship among pathways and biological processes that significantly differentiate nicotine dependent individuals from non-dependent individuals.