12

Using Ontologies in Hierarchical Modeling of Genes and Exposure in Biological Pathways

David V. Conti, Juan Pablo Lewinger, Rachel F. Tyndale, Neal L. Benowitz, Gary E. Swan, and Paul D. Thomas

Existing studies of genetic associations with nicotine dependence frequently do not reflect complex relationships between genetic, environmental, and social factors underlying tobacco use. Moreover, the scope of potential genetic variations and their impact on analysis pose a conceptual challenge to effective studies of genetic factors.

This chapter examines the potential for the use of hierarchical modeling techniques within the framework of an ontology that quantifies relationships across genotypes and phenotypes for nicotine dependence. Topics discussed include

- *An overview of the existing statistical approaches for genetic association studies in tobacco use*
- *Design and analysis considerations in the use of hierarchical modeling in conjunction with stochastic variable selection for future genetic studies of tobacco use*
- *The use of ontologies for codifying prior knowledge to support efficient computational analysis of such hierarchical models*
- *Results of a study of nicotine metabolism using the data from the Northern California Twin Registry in conjunction with the Nicotine Pharmacokinetics Ontology, showing significant genetic associations with nicotine clearance levels*

The results of this pilot study, and the potential of these approaches to overcome the methodological issues inherent in existing genetic studies, show promise for these approaches as an area for further study.

The analyses described herein were supported by National Institutes of Health grants CA084735, CA52862, DA18019, DA20830, DA02277, DA11070, and HL084705. Analysis support was also provided by University of Toronto, Canada grants CAMH, and CIHR MOP53248 and a Canada Research Chair in Pharmacogenetics.

Introduction

This chapter examines the use of ontologies as a framework for creating hierarchical models that could support quantitative, computationally driven research in biological pathways for nicotine dependence, as a potential means of linking genetic and environmental factors to yield a more accurate understanding of why people smoke. It highlights a specific example using data from the Northern California Twin Registry[1,2] to explore the heritability of nicotine metabolism, together with a discussion of broader issues involved in creating hierarchical models in conjunction with ontologies that quantify prior knowledge of relationships linking specific genotypes, endophenotypes, and phenotypes for nicotine dependence.

The multistep nature of tobacco use progression—from initiation, to episodic use, to dependence—provides several opportunities for risk factors to act. Although distinct factors may affect each step, universal factors may also create background characteristics for an individual throughout use progression. In addition, compounding the background profile are large, punctuated events, such as intervention programs, that may substantially alter an individual's tobacco use—both in isolation and synergistically with other factors. That is, smoking behavior is a composite consisting of large social factors, interpersonal relationships, and intrapersonal characteristics. Large social patterns substantially influence smoking behavior through demographic changes, financial mechanisms, cigarette availability, and perceptions of smoking. Economic factors such as unemployment rates, income levels, and cigarette prices also affect individuals' ability to purchase cigarettes.[3]

Although these large social forces often affect an individual's tobacco use, close interpersonal relationships have considerable influence as well. Personal relationships with family, friends, peer groups, and classmates form immediate surroundings and an individual's attitudes.[4] Especially among adolescents, it is within these social networks that individuals make behavioral choices about tobacco use—choices that depend on individuals' dispositional attributes as influenced by further biological, cognitive, and emotional characteristics such as the personality traits of hostility and depression.[5] Of course, these personality traits are also under some genetic influence. For example, monoamine oxidase (MAO), a mitochondrial enzyme consisting of two isoforms, MAOA and MAOB, is found in neuronal and nonneuronal cells in the brain.[6] Its main function is the breakdown of neurotransmitters; it is therefore a key enzyme in the regulation of serotonin and dopamine levels in the brain. Mouse models indicate that genetic variation within MAO are associated with changes in the levels of serotonin and dopamine in the brain and a change in behavior, especially indicators of hostility and depression.[7–9] Once an individual first smokes, the response to a particular acute or chronic dose of nicotine is determined in part by the rate of nicotine metabolism and the genes that influence metabolic function. There is a long-term physiological and psychological response as well.

To simply test the hypothesis of an association between a single genetic variant and whether an individual currently smokes or not ignores any knowledge one might have of the underlying etiologic mechanism within the analysis framework. But how does one incorporate biological information into the analysis? Often, the inclusion of previous knowledge of the underlying biological mechanism has been limited to the design phase of a study, only informing the selection of potential candidate genes and exposures. Thus, the analysis is confined to determining the independent association of each gene via contingency tables or regression models. If joint effects are suspected, the analysis is expanded to include the search for statistical interactions or, more specifically, the search for departures of independent and additive effects on the assumed scale of the outcome.[10–12] However, it is often unclear how suspected or known mechanistic and biological joint action will manifest in population-based inferences relying on epidemiological data. This chapter will examine how appropriate hierarchical models can move beyond the existing approaches to help form a framework for examining such interactions at a more macro level, which, in turn, may help to better understand and describe the role of biology in human smoking behavior.

Methodological Issues

When examining factors in nicotine dependence, difficulties in estimating and testing effects are compounded with the expanding numbers of exposures, genotypes, intermediate measures, and multiple phenotypes now readily available and relatively inexpensive to obtain on population samples with modern technologies. Such an extent of available information may lead an investigator to search for interaction effects in finer strata with limited information from the data or to exclude potentially valuable measures. For instance, possible "omic" measures such as metabolomics (e.g., surrogate measure of metabolite concentrations within a pathway),[13,14] proteomics (e.g., surrogate measures of enzyme concentration or activity within a pathway),[15,16] epigenomics (e.g., DNA methylation),[17] and interactome (e.g., protein-protein binding interactions)[18–20] are ignored in conventional gene-disease association analyses, or they are treated as an outcome in gene to intermediate-phenotype studies. Inclusion of all measures (e.g., exposures, genes, intermediate phenotypes, and disease) in a structured joint analysis may provide valuable information in clarifying the separate component contributions, their aggregate effects in complex pathways, and ultimately, determining an individual's overall risk of disease. Furthermore, each factor may contribute only a small effect that may be detected only when all relevant factors are considered together. Here, conventional regression models often reach their limits in attempting to model all these factors jointly.[21]

These difficulties have led to the development of many data-mining statistical

Ontologies: A Definition

An ontology is a formal structuring of knowledge.[a] For the purposes here, an ontology is a *formal model* of a domain of knowledge and consists of *entities* and *relations* between entities. An entity is simply a class, or category, of things that one wishes to model. An entity can be either a *continuant* (an object existing at a particular point in time) or an *occurrent* (an event or process occurring over time). Relations can be of many types, depending on the knowledge domain being represented. Two of the most common are the "is_a" relation, which specifies one class as a subclass of another (e.g., human is_a mammal), and the "part_of" relation (e.g., finger part_of hand). Ontologies have their origins in Aristotelian philosophy, but computer science has driven a renaissance in ontology development and use—initially, by the problem of representing computational knowledge in the artificial intelligence field, and subsequently, the Semantic Web.

[a]Smith, B., W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. 2005. Relations in biomedical ontologies. *Genome Biology* 6 (5): R46.

techniques aimed at detecting higher-order interactions.[22] Such approaches include tree-based methods based on recursive partitioning of the data, such as random forests[23] and logic regression.[24] For these methods, the data are split by a single binary variable into two subsamples with varying trait or outcome characteristics. These subsamples are then investigated for further splits that may be warranted on the basis of additional variables. Higher-order interactions are inferred by identifying the combination of variables and the corresponding splits that identify particular subgroups. To avoid overfitting of the data or finding splits that may exist only by chance, pruning techniques are applied to reduce the number of splits according to some pruning criteria. While these techniques can be effective at identifying higher-order interactions, there are some limitations as far as interpretability and flexibility in the modeling (e.g., including covariates and forcing in certain effects). As an alternative within the regression framework, Millstein and colleagues proposed a method called the "focused interaction testing framework."[25] This approach tests main effects and interactions among multiple candidate genes by using a series of orthogonal tests in a staged manner. Specifically, this approach tests the main effect of each candidate gene in stage 1, followed by models with two-way interactions in stage 2, and with three-way interactions in stage 3. An algorithm based on controlling false discovery rates (FDRs)[26–30] is used to control the experiment-wise type I error to a predefined level (e.g., 0.05). While simulation work has shown promise for this method when a single polymorphism is present within a gene, it is not clear how this would work when multiple correlated single nucleotide polymorphisms (SNPs) are included for several genes.

Data-mining techniques rely solely on the data for inference and ignore any prior knowledge that may exist regarding the factors of interest, specifically that these factors may be part of a biological pathway. An editorial in *Cancer Epidemiology, Biomarkers & Prevention* in 2005[31] laid out the case for pathway-driven research in molecular epidemiology and the need for further methods development in support of such research. The editorial described two broad types of approaches: one based on mechanistic modeling of specific pathways of interest, the other based on empirical modeling that incorporates what is known about the factors involved in a pathway in a flexible manner without requiring such strong parametric assumptions. The mechanistic approach can be thought of as a structural equation model in which the topology of the structure is specified by biological knowledge. This was first introduced in an application to a case-control study of colorectal polyps in relation to well-done red meat consumption, tobacco smoking, and the various genes involved in the metabolism of polycyclic aromatic hydrocarbons and heterocyclic amines that these exposures produce.[32] Here, the sequence of intermediate metabolite equilibrium concentrations was modeled in terms of a linear pharmacokinetic model, with person-specific metabolic rate parameters that depended upon their genotypes. The entire model, comprised of regression coefficients, individual- and genotype-specific population rate parameters, and their variances, was fit by using Markov chain Monte Carlo (MCMC) methods. While the authors of this chapter are continuing to investigate the statistical performance of this approach via simulation studies in a more general setting, it is recognized that the major limitation of this approach is the expertise needed to construct the topology of the mechanistic model. Unfortunately, very few biological pathways are understood well enough to specify the specific mechanism from genes to outcome. While extensions exist to estimate the topology, these methods rely heavily on accurately measured intermediates—intermediate measures that are often unmeasured in epidemiological studies or measured on only a subsample of individuals.

As an alternative to these highly parametric models, a background and extensions to hierarchical modeling are presented here. Hierarchical modeling with prior covariates aims at stabilizing and informing estimation by incorporating similarities among regression estimates using categories describing biological similarities between genes and exposures. To narrow the space of possible regression models, the prior probability of including any variable as a function of known biology is further structured. This is accomplished via Bayes model averaging using stochastic variable selection. Similar to the parametric models, these hierarchical models utilize prior knowledge and information to aid inference. However, in contrast to the highly specified mechanistic models, the knowledge only specifies exchangeable classes or sets of factors with similarities. Often this reduces to a series of indicator variables based on expert opinion. Because these opinions may be susceptible to subjective influences, the use of ontologies is proposed. Ontologies attempt to represent the knowledge base in a computable form to provide "a shared and common understanding of some domain that can be communicated between people and application systems."[33] Thus, ontologies attempt to transform implicit knowledge into specific and explicit relations. Here, a discussion is provided of how these relations may be incorporated into the hierarchical models to aid in model selection, inference, and interpretation of conclusions from observational studies.

Background on Statistical Approaches

Models for Multivariant Data within a Candidate Gene

Assume an investigation of G candidate gene regions for gene association and possible identification of specific causal variants for an identified outcome. Further assume that each gene is independent from the others—that is, no linkage disequilibrium (LD) or no underlying biological interaction between genes. A quantitative trait outcome is the focus, but by using a generalized linear framework and the appropriate link function, the discussed methods can be extended to other types of traits.[34] In addition to J exogenous or nongenetic covariates specified in the design or covariate matrix $\mathbf{W}$, assume that there are M_g finely spaced SNPs within each gene, g, and that for each polymorphism, genotype-level information for all individuals in the study is obtained. β_{gm} is used as an estimate of the risk from SNP m in gene g. For clarity, the modeling on SNPs is mainly discussed; however, the following analyses are also applicable to the modeling of other genetic markers such as microsatellites. First, treating the SNPs as independent, $\sum_g M_g$ separate regressions are performed, assuming a disease model of the form (1) where X_{gm} indicates the number of variant alleles for SNP m (i.e., additive coding) in gene g (e.g., $X_{gm} = 1$ if heterozygous and $X_{gm} = 2$ if an individual carries two copies of the variant allele), although one may also consider dominant or recessive genetic models.[35] The parameter estimate δ_{jgm} corresponds to the effects of the jth covariate on the outcome conditioned on SNP m in gene g.

$$Y \mid X_{gm}, \mathbf{W} = \beta_{0gm} + \beta_{gm} X_{gm} + \sum_{j=1}^{J} \delta_{jgm} W_j \text{ for all } m = (1,\ldots,M_g) \text{ in all } g = (1,\ldots,G) \tag{1}$$

Additional information may be incorporated if one accounts for the effect of an SNP conditional on all other SNPs within the candidate region. This is accomplished with a joint main effects model of the form (2).

Here, one sums over all the SNPs within a gene g but treat each gene as independent. The parameter estimate δ_{gj} corresponds to the effects of the jth covariate on the outcome conditioned on the SNPs in gene g. By accounting for the correlations between SNPs, this model may be useful in determining the independent contribution of each SNP within a given region, but it ignores any effects due to the arrangement of SNPs either on the same chromosome (i.e., haplotypes) or combinations of SNPs within an individual (i.e., interaction).

Aiming to capture synergistic effects between SNPs within a single candidate gene, the model in equation (2) may be extended to incorporate all interaction terms between SNPs. This model builds on the joint model in equation (2), with the form (3) where $X_{gm*g\ell} = X_{gm} * X_{g\ell}$ and "…" indicates potential higher-order interaction terms. Here, the focus is only on all pairwise second-order interactions within a gene, although one may expand this model to higher-order interactions.

In the above models, a test of the statistical significance for association to disease for each SNP can be obtained via a Wald test, score test, or likelihood ratio test (LRT) of each β_{gm}. In addition, for the main effects model (2) and interaction model (3), one may perform an omnibus LRT comparing a full model in which β_{gm} is estimated for each of M_g markers, $\hat{\beta}_g = \left(\hat{\beta}_{gm}, \ldots, \hat{\beta}_{gM_g}\right)$, to the null model in which all SNP effects within the gene are set to zero. This global M_g-degree of freedom LRT provides evidence for an overall association of the chromosomal segment to disease.

When multiple SNPs are available within a gene, an alternative is to analyze the association of haplotypes to disease. For a given set of haplotypes, H_g, the haplotypic risk may be modeled by using a similar logistic regression for H_g–1 of the haplotypes (4).

Here, X_{gh} is used as an indicator variable denoting the number of haplotypes of type h that an individual possesses within gene g. Usually for the haplotype model, the most common haplotype, h_{g1}, acts as the referent haplotype. Similar to the SNP analysis, a Wald test, score test, or LRT statistic may be calculated for each η_{gh} to test associations with each haplotype. In addition, an omnibus LRT can test the overall association of the gene region to the trait. If haplotypes are unknown, one may substitute for X_{gh} an expected probability for haplotype h.[36,37] Haplotype-based analysis as outlined in equation (4) has been advocated because of the potential reduction in the number of comparisons made (because there are usually fewer common haplotypes than common SNPs), the ability of haplotypes to better exploit patterns of LD, and the capacity to capture causal effects that may

$$Y \mid X_{g1}, \ldots, X_{gM_g}, \mathbf{W} = \beta_{0g} + \sum_{m=1}^{M_g} \beta_{gm} X_{gm} + \sum_{j=1}^{J} \delta_{gj} W_j \text{ for all } g = (1, \ldots, G) \tag{2}$$

$$Y \mid X_{g1}, \ldots, X_{gM}, \mathbf{W} = \beta_{0g} + \sum_{m=1}^{M_g} \beta_{gm} X_{gm} + \sum_{m=1}^{M_g} \sum_{\ell \neq m}^{M_g} \beta_{gm*g\ell} X_{gm*g\ell} + \ldots + \sum_{j=1}^{J} \delta_{gj} W_j \text{ for all } g = (1, \ldots, G) \tag{3}$$

$$Y \mid X_{g1}, \ldots, X_{gH}, \mathbf{W} = \beta_{0g} + \sum_{h=2}^{H_g} \eta_{gh} X_{gh} + \sum_{j=1}^{J} \delta_{gj} W_j \text{ for all } g = (1, \ldots, G) \tag{4}$$

be due to a combination of variants on the same chromosome.[38,39] However, to attain these potential benefits one must often narrow each region to identify a limited number of haplotypes; this is typically done by identifying blocks or continuous regions of high LD along the chromosome. This, in turn, makes haplotype analysis subject to how one determines these regions via the underlying LD structure and the accompanying uncertainty in that determination.[40–43]

As an alternative to haplotype analysis, Conti and Gauderman[44] proposed a modified pairwise interaction term to capture phase information in equation (3) to allow for most of the haplotype information in the data to be exploited, without having to consider all possible haplotype resolutions, as required for equation (4). At the genotype level within gene *g,* one can approximate haplotype information by modifying the second-order interaction terms in model (3) to describe the phase between pairwise SNPs, m and ℓ, and given the two haplotypes for individual *i,* h_{ig1} and h_{ig2}. Specifically, the definition is given in equation (5).

The above coding assumes that the *cis* configuration or double variant haplotype is additive to disease. However, it is also possible that the *trans* configuration of the variant alleles, as defined here, may be at higher risk. In this alternative case, one can specify the reverse coding for the double heterozygotes (i.e., if $X_{gm}*X_{g\ell} = 1$, and h_{ig1} or h_{ig2} is the double variant haplotype, then $X_{gm.g\ell} = 1$). This parameterization allows for separate tests for each SNP effect *(β_{gm}),* pairwise phase term *($\beta_{gm.g\ell}$),* and the overall contribution of the candidate region to disease via a global LRT. When the phase is unknown, the *cis* phase term is altered to reflect the probability of a *cis* haplotype in the population for each pair of loci assuming Hardy-Weinberg equilibrium. As an example, assume two SNPs, **A** and **B,** each with two alleles, *(A, a)* and *(B, b),* respectively, as well as the four possible haplotypes, *AB, Ab, aB,* and *ab*. Thus, one can calculate the probability of the *cis* configuration of the two SNP haplotypes as given in equation (6) where $P(AB)$, $P(Ab)$, $P(aB)$, and $P(ab)$ are estimated from genotype data using the expectation maximization, or EM, algorithm.[45] This is equivalent to altering the phase term in equation (5) by setting $X_{gm.g\ell} = \rho^{(cis)}_{gm.g\ell}$ if $X_{gm}*X_{g\ell} = 1$. Thus, a genotype model with phase interaction terms not only avoids long-range haplotype estimation but also allows for the investigation of which SNPs are driving the association within each candidate gene. In addition, this model provides a flexible framework for incorporating relations among numerous SNPs over several candidate genes.

Extensions to Multiple Genes and Exposures

The above models present various alternatives to the analysis of numerous variants within a candidate gene, with the

$$X_{gm.g\ell} = \begin{cases} 2 & \text{if } X_{gm} * X_{g\ell} = 4 \\ 1 & \text{if } X_{gm} * X_{g\ell} = 2 \\ 1 & \text{if } X_{gm} * X_{g\ell} = 1, \text{ and } h_{ig1} \text{ or } h_{ig2} \text{ is a double variant haplotype} \\ 0 & \text{if } X_{gm} * X_{g\ell} = 1, \text{ and } h_{ig1} \text{ and } h_{ig2} \text{ is } \textit{not} \text{ a double variant haplotype} \\ 0 & \text{if } X_{gm} * X_{g\ell} = 0 \end{cases} \tag{5}$$

$$\rho^{(cis)}_{A.B} = \Pr(AB, ab) = \frac{P(AB)P(ab)}{P(AB)P(ab) + P(Ab)P(aB)} \tag{6}$$

increase in complexity aiming to better capture the LD and joint effects of multiple SNPs. The complexity may be warranted if a true causal SNP is not measured, and the analysis must rely on how the combination of measured SNPs captures the underlying effect. Of course, if a true causal variant(s) is measured, the most appropriate model may be the one that focuses solely on that variant(s), ignoring all others. In contrast, it may be the combination of several SNPs acting together that leads to variation in the outcome. In this case, simple tests of the marginal effect of each SNP may not be sufficient, and interaction terms may be necessary to detect these higher-order joint actions. Thus, even within a single gene, there are uncertainties regarding the most appropriate model to use. These uncertainties only increase as one attempts to evaluate multiple candidate genes, each with multiple polymorphisms. The previous models treat each candidate gene as independent. This assumption may be adequate if the genes are unlinked, and therefore, SNPs between candidate genes are not in LD. However, because a set of candidate genes is most often selected with a priori knowledge that they act via an underlying biological mechanism or pathway, there is a good possibility that interactions may be present across genes. Their linear modeling framework may be expanded to accommodate multiple gene effects as given in (7) by summing over all possible genes G and, within each gene, including all marker main effects and phase terms, and including interaction terms across genes (7).

For similar reasons, one may also want to investigate gene-environment interaction with measured covariates. This expands the model further as in (8).

The Challenge of Numerous Polymorphisms and Exposures

The investigation of associations for numerous polymorphisms within a single candidate gene and across multiple genes can raise concerns about multiple comparisons and sparse data bias in estimation. As one extreme approach, each polymorphism can be treated as independent, as in model (1). This approach is problematic: these reduced models may result in underestimated variance, and they do not account for the correlation that may exist among the polymorphisms, such as two polymorphisms in LD with each other within a gene region.[46] Furthermore, treating each polymorphism as independent and relying on statistical tests across all polymorphisms can lead to issues of multiple comparisons. While one may perform adjustment in the declaration of significance, such as a Bonferroni correction or control of the false discovery rates,[27,47,48] these procedures may not accurately account for the relations between the

$$Y \mid \mathbf{X},\mathbf{W}] = \beta_0 + \sum_{g=1}^{G}\left(\sum_{m=1}^{M_g}\beta_{gm}X_{gm} + \sum_{m=1}^{M_g}\sum_{\ell \neq m}^{M_g}\beta_{gm.g\ell}X_{gm.g\ell}\right) + \left(\sum_{g=1}^{G}\sum_{m=1}^{M_g}\sum_{k \neq g}^{G}\sum_{\ell=1}^{M_k}\beta_{gm*k\ell}X_{gm*k\ell}\right) + \sum_{j=1}^{J}\delta_j W_j \tag{7}$$

$$\begin{aligned} Y \mid \mathbf{X},\mathbf{W}] = \beta_0 &+ \sum_{g=1}^{G}\left(\sum_{m=1}^{M_g}\beta_{gm}X_{gm} + \sum_{m=1}^{M_g}\sum_{\ell \neq m}^{M_g}\beta_{gm.g\ell}X_{gm.g\ell}\right) \\ &+ \left(\sum_{g=1}^{G}\sum_{m=1}^{M_g}\sum_{k \neq g}^{G}\sum_{\ell=1}^{M_k}\beta_{gm*k\ell}X_{gm*k\ell}\right) \\ &+ \sum_{j=1}^{J}\delta_j W_j + \left(\sum_{g=1}^{G}\sum_{j=1}^{J}\delta_{gj}X_g W_j\right) \end{aligned} \tag{8}$$

polymorphisms, and they do not yield estimates of effect conditional upon other polymorphisms and exposures.

At the other extreme, the analyst may choose to model all main effects and interactions in one single model, as described in equation (8). Including all genetic polymorphisms and exposures in one model can lead to biased and unreliable estimates due to sparse data when the number of parameters approaches the number of individuals in the sample.[21,49] These models tend to overfit the data, resulting in estimates that explain the observed data well, but will lead to unrealistic predictions for any new data or biased inferences implied by the estimates. While conceptually attractive, in modern observational studies this approach quickly reaches the limits of the data, especially given the relatively large expense of enrolling an individual into a study in comparison to the rapidly dropping costs of obtaining a plethora of genotype-level information for a given individual. Often, a compromise in analysis approaches comes in the form of model selection or using the data and/or prior information to determine which set of polymorphisms and exposures may have substantial effects and only include those terms in the model. Models (1) through (7) may be viewed as types of reduced models in which polymorphisms and/or genes are assumed to be independent or interacting effects are assumed to be nonexistent. The use of knowledge or statistical tests is attractive in providing the analyst with simplified models in which to estimate and interpret. However, it is important to realize that, by not including a certain term in the model, the analyst is implicitly stating a belief that, with 100% certainty, that term's effect estimate is zero. Is previous knowledge reliable enough to justify the exclusion of a term, or is there a level of uncertainty? Clearly, relying solely on a priori decisions of what to include in the model is limited to the accuracy of the prior knowledge and, moreover, these a priori decisions ignore the data completely. In contrast, model selection procedures that use only the data to decide which terms to include in the model may underestimate the variance for each term by not accounting for the uncertainty in the selection procedure itself. Furthermore, automated procedures are prone to increased type I errors (i.e., false positive errors) by relying strictly on statistical cutoffs in the model-building process.[50]

Potential Solutions

To address these issues, an approach is proposed that uses hierarchical modeling and stochastic variable selection. Hierarchical modeling allows for the construction of complex probability models that incorporate higher-level information to yield more stable and plausible measures of association. Stochastic variable selection utilizes both the data and prior knowledge to determine which terms to include in the models, resulting in a guided model search leading to more representative and interpretable models. These approaches are possible because the hierarchical nature of the data—that is, polymorphisms within genes and genes within pathways—provides an opportunity to formalize a bottom-up approach placing more emphasis on combinations of polymorphisms within a gene in comparison to combinations across genes. This hierarchy served as the foundation for the development of the various approaches outlined in equations (1) through (8). This culminates in the saturated model (8) in which one first sums over SNPs main effects and SNP interaction effects within a gene, then SNP interaction effects across genes, and finally, over SNP $\times$ covariate interactions. It is proposed to formalize the combination of knowledge-based heuristics and model selection procedures in deciding which model is most appropriate. In this context, hierarchical modeling and stochastic variable selection as conventionally applied are briefly introduced, and then these approaches are combined in a pathway-based model.

Hierarchical Modeling

A primary motivation for using hierarchical modeling and stochastic variable selection with structured priors is to describe the joint distribution of the underlying genetic structure and biological mechanism represented by the data and, notably, the uncertainty in representing that structure and mechanism. In doing so, the parameter estimates and corresponding uncertainty intervals will better capture the dependency between terms; therefore, the resultant tests will more effectively reflect the evaluation of multiple polymorphisms and exposures.[49–53] This is similar in spirit to an approach proposed by Wacholder and colleagues[54] in which they introduced a Bayesian procedure for multiple comparisons that incorporates a prior specification of the probability of any given polymorphism being associated with an outcome. While the notion of incorporating prior knowledge into testing and estimation is an important one, the false positive reporting probability of Wacholder and colleagues[54] frames the decision into a binary choice between the null hypothesis and an effect size determined from estimation using observed data.[55] Alternatively, one can specify a prior distribution for the effect size via a hierarchical model. By incorporating known information regarding the relations among the genetic polymorphisms, a joint distribution is specified that both stabilizes the final effect estimates and incorporates dependencies across multiple tests of association. Specifically, one can model the regression coefficients β_{gm} from model (9) in terms of a regression on a vector of ρ "prior covariates" $\boldsymbol{Z_{gm}} = (Z_{g1},\ldots Z_{g\rho})$. Thus, a second-level model of the form is adopted (9).

$$\begin{aligned} \underline{\beta} &= \mathbf{Z}_{\mu}\,\underline{\mu} + \underline{U} \\ \underline{U} &\sim N(0, \tau^2\Sigma) \end{aligned} \tag{9}$$

The design matrix $\mathbf{Z}_{\mu}$ contains the second-stage covariates reflecting higher-level relations between the polymorphisms, $\underline{\mu}$ is a column vector of coefficients corresponding to these higher-level effects on the trait outcome, $\underline{U}$ is a column vector of random effects capturing the residual variation after adjustment by the relations in $\mathbf{Z}_{\mu}$, and Σ is a covariance matrix specifying any residual covariance among the regression coefficients. This hierarchy results in posterior estimates of effect for the polymorphisms $\tilde{\beta}$, which are an inverse-variance weighted average between the maximum likelihood estimates obtained from a conventional regression and the estimated conditional second-stage means $\mathbf{Z}_{\mu}\,\underline{\mu}$. Thus, the final estimates of effect are dependent upon the amount of information available. Estimates with less information may be unstable and will tend to have larger variances. This, in turn, will result in a final posterior estimate more heavily weighted toward the prior information reflected by the conditional second-stage means $\mathbf{Z}_{\mu}\,\underline{\mu}$.

An important assumption here is that the modeled parameters, the βs, are exchangeable. Formally, this means that conditional on the information in $\mathbf{Z}_{\mu}$, the parameters have no prior ordering or grouping such that their joint distribution, $f(\beta_{11},\ldots,\beta_{gm})$, is invariant to permutations of the indexes $g = (1,\ldots,G)$ and $m = (1,\ldots,M)$. If this assumption holds, one may assume that the parameters are drawn from the same population distribution. In practice, the validity of this exchangeability assumption requires one to both focus on the interpretation of the βs and on how to group them via the design-matrix $\mathbf{Z}_{\mu}$. First, in linear regression, the βs represent the increase in the outcome, Y, given a one-unit increase in the independent variable, X. In the analysis of SNPs, these effect estimates are the increase in Y given one unit in change in the variant allele, assuming an additive coding for a selected variant. Here, across numerous SNPs, all the βs reflect similar interpretations for their corresponding effect estimates, and the design matrix $\mathbf{Z}_\mu$ may be constructed to incorporate a priori knowledge of SNPs having similar estimates of effect, for example. However, even in this simple example of multiple SNPs, care must be taken in how the effect estimates are interpreted and the impact this interpretation will have on the construction of the prior covariates. Specifically, for a particular SNP, one must consider which of the two alleles describes the increase in effect.[56] If it is known that two SNPs might share similar risks—that is, are exchangeable classes conditional on the design matrix—one is really assuming that the two variant alleles as defined at the two respective SNPs share similar effects *in the same direction.* If one does not have knowledge of the direction of effect for each variant, then one may incorrectly specify sharing of two effects that act in opposite directions and are thus not from the same population distribution.[57] Further complications arise as the analysis is expanded to include environmental covariates. On what scale does one define the effect estimates corresponding to environmental covariates? And, is the corresponding effect estimate exchangeable with other covariate or SNP estimates? To address these difficulties, it is proposed that the hierarchical model be modified to model the test statistics rather than the effect estimates. This will be discussed later in the "Methods" section.

Model Selection via Stochastic Variable Selection

Although hierarchical modeling can stabilize the estimates of effect across the numerous terms, it is also of interest to highlight the linear combinations of SNPs and phase terms that best capture the gene-disease relations. Furthermore, it is desirable to account for varying prior beliefs that each polymorphism or term is involved in the trait outcome. That is, although all the genes were chosen with at least some belief that they are involved in the outcome under investigation, some genes are more likely to be involved given prior functional evidence or knowledge of the underlying biology. Similarly, within a specific gene, some polymorphisms are more likely to affect trait variation, with some polymorphisms chosen because of putative functional evidence and others chosen strictly to capture an unknown causal effect via LD. Thus, a stochastic variable selection approach is proposed to stochastically search the model space to highlight important SNP and phase terms and to average over all possible models. This approach has the advantage of accounting for uncertainty in model selection and allowing for a flexible prior structure in which one can incorporate the relations among terms when selecting representative models. Previously,[58] a variation was implemented of the stochastic search variable selection algorithm, first presented by George and McCulloch,[59] by introducing a latent variable, $\gamma_v = 0$ or 1, indicating whether a term, *v,* is included in a model (10).

The above specification is conventionally implemented with a prior second-stage normal distribution with a mean of zero. While others have discussed the use of semi-Bayes or empirical-Bayes approaches to prespecify ψ and σ^2,[59–61] a fully Bayesian approach can be adopted to integrate over posterior distributions using MCMC methods as implemented in WinBUGS.[44,58,62]

$$\Pr(\beta_v \mid \gamma_v, \psi_t, \sigma) = \begin{cases} 0 & \text{if } \gamma_v = 0 \\ N(0, (\psi_t \sigma)^2) & \text{if } \gamma_v = 1 \end{cases} \tag{10}$$

The posterior probability of $\gamma = 1$ and the set of possible models visited during the stochastic search can be used to gauge the impact of each term and the combinations of SNPs and phase terms that best explain the relation of genetic variation to disease. These posterior probabilities will depend on the specified prior distribution for γ. The simplest form of the prior is to assume a binomial prior distribution for γ (11) where ρ_v is the probability that $\gamma_v = 1$. By assuming that ρ_v is constant for all terms, one assumes that the corresponding parameters, β_v, are exchangeable, as indicated in equation (10), and equally likely to be included in any given model. Specifically, main effects and interaction terms are equally likely. However, it is desirable to structure the prior in equation (11) to both guide the stochastic search to models that are more parsimonious in relation to the number of SNPs or terms included in any given model, and also to assist the stochastic search in the inclusion of phase terms, conditional on the inclusion of both "parent" SNPs used to define the pairwise interaction term.[63] Following Conti and Gauderman,[44] the level of parsimony can be controlled by setting the prior for ρ for SNP main effects as $\Pr(\rho_{SNPs}) = \text{Beta}(1,3)$. This gives a low expected probability ($E[\rho_{SNPs}] = 0.25$) of inclusion for any given SNP and places emphasis on models with fewer SNPs.

Furthermore, following Chipman,[63] a conditional probability for inclusion of phase terms, $\gamma_{gm.g\ell}$, is defined as (12). This conditional prior reduces the model space visited by the stochastic search. This structure is invoked to reflect the approximation of the underlying haplotype architecture with linear combinations of SNP and, if necessary, phase terms. Introducing a hierarchical dependency of phase terms on the "parent" SNP terms directs the stochastic search to simpler and more stable models, if appropriate. To offset this restriction and to encourage the exploration of the importance of phase, a higher probability is specified for the inclusion of phase terms, conditional on the inclusion of both "parent" SNPs, $\Pr(\rho_{gm.g\ell} \mid \gamma_{gm} = 1, \gamma_{g\ell} = 1) = \text{Beta}(3,1)$. This puts a higher prior expected probability on the phase terms ($E[\rho_{gm.g\ell} \mid \gamma_{gm} = 1, \gamma_{g\ell} = 1] = 0.75$). However, marginally, the prior expected probability for the inclusion of any given phase term is lower than the SNP main effects, $E[\rho_{gm.g\ell}] = 0.05$. This structured prior in equation (12) acts to both guide the stochastic search to models that are more parsimonious in relation to the number of SNPs included in any given model, and also assists the stochastic search in the inclusion of interaction terms, conditional on the inclusion of both "parent" main effects used to define the interaction term. In a similar fashion, the structured priors can be used to limit and/or guide the model search to combinations of SNPs across genes within subpathways and networks, and models can thus be summarized.[64]

Methods

General regression approaches to the analysis of multiple SNPs within and across candidate genes have been reviewed. Also, both hierarchical and model selection extensions to these regression models have been discussed. As mentioned earlier,

$$f(\gamma_1,\ldots,\gamma_V \mid \rho_1,\ldots,\rho_V) = \prod_{v=1}^{V} \rho_v^{\gamma_v}(1-\rho_v)^{(1-\gamma_v)} \tag{11}$$

$$\Pr(\gamma_{gm.g\ell} = 1 \mid \gamma_{gm}, \gamma_{g\ell}) = \begin{cases} \rho_{gm.g\ell} & \text{if } (\gamma_{gm}, \gamma_{g\ell}) = (1,1) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

it is desirable to combine the benefits of borrowing information from exchangeable classes via hierarchical modeling with the ability to search the expansive model space via stochastic variable selection. However, the combination of these two approaches introduces two notable practical difficulties. First, how does one reasonably define exchangeable classes across SNPs (and possibly microsatellites), environmental factors, and all possible interaction terms? Second, how does one search such a vast space of applicable models in a reasonable amount of computational time? One can begin by first framing the hierarchical model on unsigned summary statistics from the regression model rather than from the effect estimates. Thus, one only has to define exchangeable classes for the unsigned summary statistics and avoid specification for the effect estimates, which may vary in scale and direction. In addition, an empirical Bayes approach is used to regress these summary statistics on prior covariates to yield posterior estimates of the summary statistics and of the probability that a summary statistic is nonzero. This probability, in turn, determines the probability that a corresponding term is included in the model.

Hierarchical Model

Following Lewinger and colleagues,[65] one can begin by defining a Wald test statistic, $T_v^2 = \hat{\beta}^2 \big/ \hat{\mathrm{var}}(\hat{\beta})$, for a specific term, v, in a regression model. This test statistic has an asymptotic χ^2 distribution with one degree of freedom and a noncentrality parameter λ_v^2. Since the interest is in defining exchangeable classes independent of the direction of effect, the focus is on the unsigned statistic $t_v = +\sqrt{T_v^2}$ and the corresponding noncentrality parameter $\lambda_v = +\sqrt{\lambda_v^2}$, resulting in an asymptotic distribution for t_v as $\chi_1(\lambda_v) = |N(\lambda_v,1)|$. Specifically, this distribution is of the form (13) where φ is the standard normal density. This places a second-stage distribution on the test statistics obtained from a first-stage regression model. Of interest is deciding if this test statistic provides evidence for the SNP or factor being involved in the outcome of interest. If $\lambda_v = 0$, there is no association with the outcome. Positive values for λ_v indicate an association with increasing evidence as λ_v grows in magnitude. This can be formalized by modeling the λs as a mixture between a chi distribution with a positive noncentrality parameter and a point mass $\delta(0)$ where $\lambda_v = 0$ (14).

Here, H_v is an indicator of whether a term is associated with the outcome and p_v is the corresponding probability of that association. Given that there is a true association, the expected noncentrality parameters are influenced by e_v and $\sigma > 0$.

The terms in a regression model are not all equivalent in regard to prior information that may be available (e.g., is the SNP known or predicted to affect gene function? Or, how important is the gene in a particular pathway?) and in regard to the hierarchical structure of the model (i.e., main effects versus interactions). Recognizing that differences exist across terms, both the probability that the association is true and magnitude of the noncentrality parameter are regressed on

$$t_v \mid \lambda_v \sim \varphi(t_v - \lambda_v) + \varphi(t_v + \lambda_v),\ t_v \geq 0 \tag{13}$$

$$\lambda_v \mid H_v, e_v \sim H_v \sigma^{-1} \chi_1(\sigma^{-1} e_v) + (1 - H_v)\delta(0) \tag{14}$$
$$H_v \sim Bernoulli(p_v)$$

a set of prior covariates constructed to indicate the prior knowledge (15).

$$\begin{aligned} \operatorname{logit}(p_v) &= \underline{\pi}' \mathbf{Z}_{\pi,v} \\ e_v &= \left| \underline{\mu}' \mathbf{Z}_{\mu,v} \right| \end{aligned} \qquad (15)$$

The intercept μ_0 is constrained to be nonnegative for identifiability. For details regarding the estimation of the relevant parameters, see appendix 1.

Model Selection via Stochastic Variable Selection

The above hierarchical model incorporates prior information into the estimation of the posterior probability that a term is associated with an outcome and the magnitude of the test statistics. However, the model assumes that there is a given regression model in which to obtain the corresponding test statistics. Given the immense space of all possible models outlined in equations (1) through (9), it is desirable for the priors and the data to influence which models are examined. Following the previous discussion on stochastic variable selection, a vector of variables, $\underline{\gamma}$, is introduced that indicates if a certain term is included in the model. Conditional on the selected terms, the test statistic is then calculated as (16).

$$t_v \mid \gamma_v = \begin{cases} \dfrac{\hat{\beta}_v^2}{\hat{\operatorname{var}}(\hat{\beta}_v)} & \text{if } \gamma_v = 1 \\ \chi_1(0) & \text{if } \gamma_v = 0 \end{cases} \qquad (16)$$

To allow for both the priors and the data to influence model selection, one sets the probability that a term is included in a regression model to be equal to the probability that an association is true, that is, (17).

$$\gamma_v \sim Bernoulli(p_v) \qquad (17)$$

Because of the hierarchical nature of the terms within a given regression model, a similar conditioning as outlined in equation (12) for the inclusion of interaction terms is included. For more details regarding the model selection algorithm, see appendix 2.

The prior structure specified via $\mathbf{Z}_\mu$ and $\mathbf{Z}_\pi$ and incorporated into this hierarchical model serves two purposes. First, it allows the posterior estimates of $\hat{P}_v$ and $\hat{E}_v$ to borrow information from exchangeable classes, and second, via $\hat{P}_v$, it will guide the stochastic search to regression models that include more biologically relevant terms. The overall impact of these structured priors is to narrow the space of possible models searched via the stochastic algorithm. Thus, instead of being faced with an impossible number of main effect and interacting terms and possible models, the process is reduced with biological knowledge to an informed and guided search procedure.

In the process of the stochastic search, the data will serve to update the prior probability and inform one of the impact of each factor via the posterior estimates for $\hat{P}_v$, $\hat{E}_v$, and the posterior probability of certain terms being selected. For inference regarding the posterior magnitude of the test statistic, calculation (18) is made. For the posterior probability of an association to be true, calculation (19) is made.

Because the final inference regarding the importance of each factor via the posterior probability of association and the probability of each factor being selected must reflect the prior probability structure, one relies on Bayes factors for inference.[66] Bayes factors are the ratio of the posterior probability odds, comparing two hypotheses to the prior probability odds, and can be thought of as a type of marginal likelihood ratio for the comparison of two hypotheses. For example, calculate as in equation (20).

$$\tilde{E}_v = E[E_v \mid \gamma_v = 1, \mathbf{Y}, \mathbf{X}, \mathbf{Z}_\mu, \mathbf{Z}_\pi] = \frac{1}{N_v} \sum_{i=1}^{N_v} \hat{E}_v^i \tag{18}$$

$$\tilde{P}_v = E[P_v \mid \gamma_v = 1, \mathbf{Y}, \mathbf{X}, \mathbf{Z}_\mu, \mathbf{Z}_\pi] = \frac{1}{N_v} \sum_{i=1}^{N_v} \hat{P}_v^i \tag{19}$$

$$BF(\gamma_v) = \frac{\Pr(\gamma_v = 1 \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}_\mu, \mathbf{Z}_\pi) \Big/ \left(1 - \Pr(\gamma_v = 1 \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}_\mu, \mathbf{Z}_\pi)\right)}{\Pr(\gamma_v = 1 \mid \mathbf{X}, \mathbf{Z}_\mu, \mathbf{Z}_\pi) \Big/ \left(1 - \Pr(\gamma_v = 1 \mid \mathbf{X}, \mathbf{Z}_\mu, \mathbf{Z}_\pi)\right)} \tag{20}$$

The numerator is calculated by using the frequency distribution of $\gamma_v = 1$ from the MCMC iterations when examining the association to **Y.** In a similar fashion, the denominator is calculated by calculating the frequency distribution of $\gamma_v = 1$ under the null of no association to **Y,** or effectively removing **Y** from the conditioning. Also, a Bayes factor should be calculated for the posterior probability of a true association. For each hypothesis comparison, a Bayes factor between 3 and 20 can be considered as "positive" evidence, 20 to 150 as "strong" evidence, and greater than 150 as "very strong."[66]

Prior Knowledge and Ontologies

The list of candidate genes has been chosen because they are involved in biological pathways suspected in the trait process. Thus, in branching out from assessing the impact of a single candidate gene to comprehensively evaluating the factors within interconnected pathways, one is faced with the a priori possibility that many interactions, often of higher order, will exist between factors (as represented in model [9]).

For genetic association studies, one wants to encode in computational form prior knowledge that can either (1) estimate the likelihood of effects from a specific genotypic or phenotypic variable or (2) hypothesize a relationship between two or more variables that would otherwise be assumed to be independent: genotype-genotype relationships, phenotype-phenotype relationships, and genotype-phenotype relationships. Knowledge of these types has been used in association studies, but only in either the study design phase or as an independent analysis, not as an integral part of a global analysis as proposed here. As a familiar example, "coding SNPs" are often prioritized in genotyping studies because they cause a change in the protein product of a gene—either a missense (amino acid substitution) or nonsense (premature stop codon) change—that is, more likely to have a phenotypic effect than a random, noncoding SNP. Interactions are often tested between polymorphisms within a particular gene because they have a relatively high likelihood of interacting simply by virtue of being in the same gene. The LD provides knowledge of haplotype structure in the population that can be used to select SNPs for genotyping.[67]

Prior Knowledge About Potential Functional Effects of Genetic Polymorphisms

A number of prediction algorithms exist for estimating the probability that a given genetic polymorphism may have phenotypic consequences. Most human polymorphisms are believed to have little or no detectable phenotypic effect;[68] it is almost certainly true that in any given genetic association study, the probability that a randomly chosen polymorphism affects the phenotype of interest is vanishingly small. A number of computational methods have been developed to estimate the probability that a polymorphism has a functional effect, such as the Sorting Intolerant From Tolerant, or SIFT, procedure;[69] PolyPhen (polymorphism phenotyping);[70] and subPSEC (substitution position-specific evolutionary conservation).[71]

Most of these methods apply to nonsynonymous coding SNPs (SNPs that result in an amino acid substitution in the protein product of a gene), and they predict the probability specifically of a deleterious effect. The most commonly used methods analyze either (1) related protein sequences and judge a polymorphism to be deleterious if it causes a substitution at a highly conserved site (because conservation is due to natural selection *against* substitutions at that site)[72–74] or (2) how the change may disrupt known elements of protein 3D structure (e.g., substitutions in the interior of proteins are more likely to destabilize protein structure).[72,75] Figure 12.1 shows examples of these analyses, applied to the **2* variant of *CYP2A6* (cytochrome P450, subfamily 2A, polypeptide 6), which changes leucine 160 to histidine (L160H). This substitution completely inactivates the enzyme. This substitution can be predicted as deleterious by using evolutionary analysis (figure 12.1A): all CYP2A (and 2B and 2G, not shown) enzymes have either leucine, isoleucine, or phenylalanine (large hydrophobic amino acids) at that position, so histidine, which is polar, would be predicted to be deleterious. A structure-based analysis (figure 12.1B) shows that position 160 is on the interior of the protein, also predicting a probable deleterious effect.

Analysis of conservation patterns can also be applied to noncoding DNA sequences,[76] although there is generally less statistical power than for coding SNP analysis. Noncoding DNA sequences generally diverge faster than protein sequences, and local mutation rates can be difficult to estimate in the absence of known, neutrally evolving sites (which in proteins can be estimated from synonymous coding changes).

It is important to note that these analyses provide prior information about the likelihood of a particular genetic change resulting in a phenotypic change, although not necessarily the phenotypic change (or changes) assayed in any particular association study.

Systems Biology and Genetic Association Studies

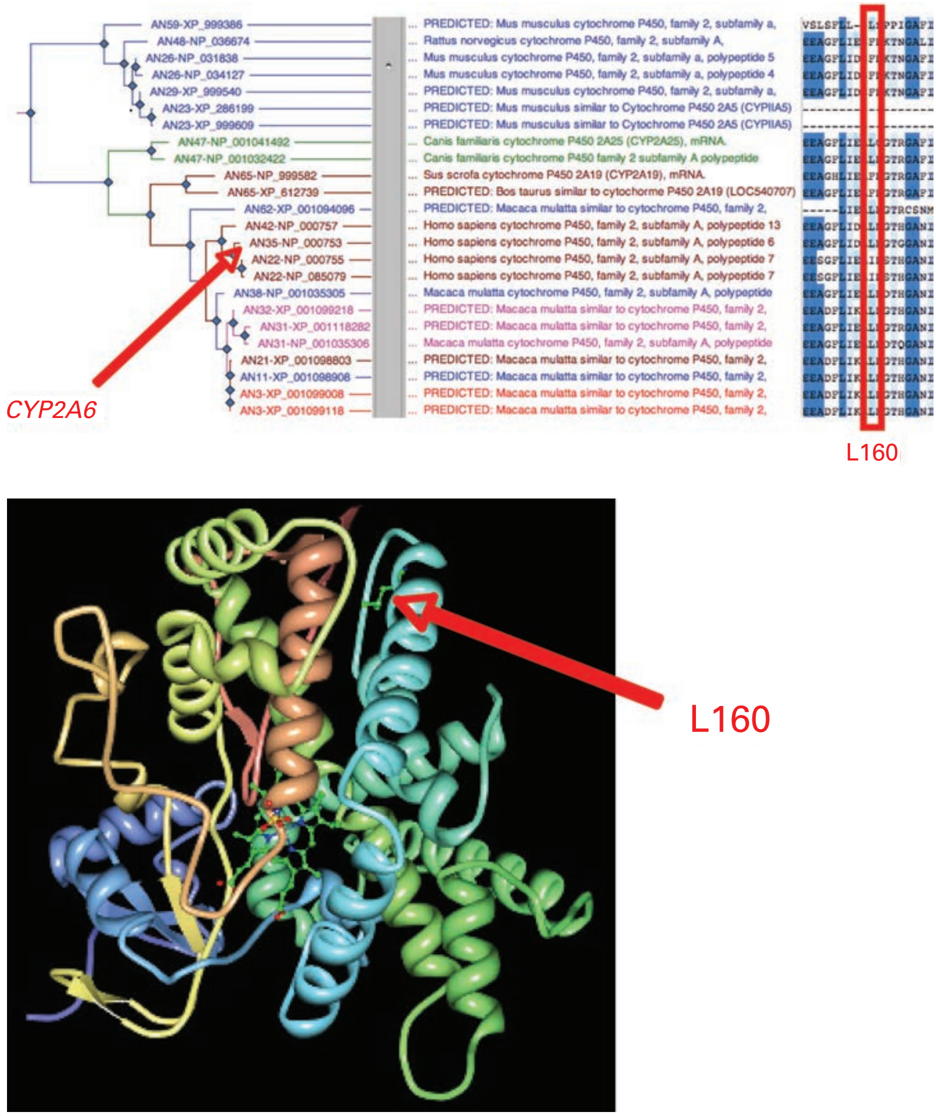
Of interest is building a model of the system (including both biological and environmental variables) that represents how a perturbation in any one variable will affect other variables in the system. The more information in the model, the more information can be used to infer effects of changes in genetic or environmental variables. In genetic association studies, the minimal set of variables includes genetic polymorphisms and phenotypic effects (outcomes). One way to visualize this system is in terms of a network model, in which nodes represent variables and edges represent potential paths for propagating perturbations to the system. Examples of networks are given in figure 12.2. In this model, a variant allele (e.g., a polymorphism) of a gene is a "perturbation" of the system relative to the wild-type allele. An association between a polymorphism and a phenotype implies that the perturbation due to the polymorphism was propagated through the system to affect the phenotype. If the phenotype is a defined event—for example, smoking cessation—then the polymorphism might affect the probability of occurrence of the event. If the phenotype is a quantitative trait—for example, cigarettes smoked per day—then the polymorphism might affect the magnitude or variance of the trait.

In the simplest case, a genetic association study will collect data only on one or

Figure 12.1 Evolutionary and Structural Analyses for the *CYP2A6*2* Variant

A.

B.

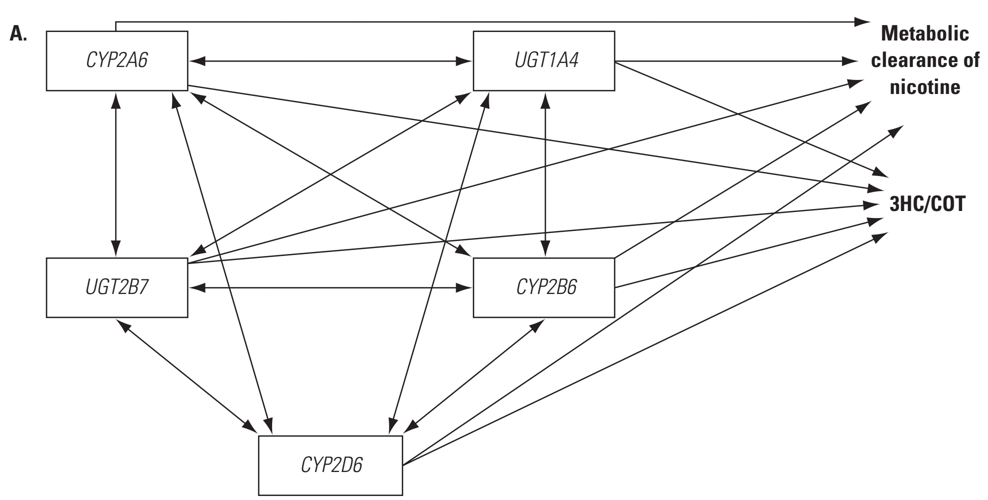


Note. In the top figure (A), evolutionary conservation in homologs of human *CYP2A6* at position 160 predicts that *CYP2A6*2* will have reduced function. Only large, hydrophobic amino acids (L:leucine, F:phenylalanine, I:isoleucine) are found in all homologues while the *CYP2A6*2* variant codes for a histidine at this position. The left panel shows a phylogenetic tree of some mammalian homologues of human *CYP2A6,* the middle panel provides information about each sequence including the gene and organism (mouse: *Mus musculus;* rat: *Rattus norvegicus;* dog: *Canis familiaris;* pig: *Sus scrofa;* cow: *Bos taurus;* macacque: *Macaca mulatta;* human), and the right panel shows part of the sequences of the corresponding protein near position 160 in *CYP2A6.* The tree was constructed using nonsynonymous divergence as the distance metric. mRNA = messenger RNA. In the bottom figure (B), protein structure analysis predicts that *CYP2A6*2* will have reduced function. Leucine 160 is buried inside the hydrophobic core of *CYP2A6,* where a substitution to histidine in *CYP2A6*2* would be predicted to destabilize the protein. Picture drawn with ProteinWorkshop.[77]

more phenotypes and one or more genetic polymorphisms. If there is no prior knowledge about the potential relationships between these genetic polymorphisms and the phenotype, one implicitly assumes a completely connected network, in which any polymorphism can affect any phenotype by any path. An example of such a network, for the genes and phenotypes considered in this paper, is shown in figure 12.2A. In this network, each polymorphism is assumed, prior to data analysis, to have an equal probability of affecting the phenotype. All interaction terms are also considered a priori to be equally probable. The model makes no assumptions about the underlying mechanisms by which genetic (or environmental) perturbations will propagate to the phenotypes of interest. In this sense, it is hypothesis free, although in practice most genetic association studies focus on "candidate genes" that are hypothesized a priori to have a potential role in a particular phenotype.

An increasing amount of information is becoming available about the underlying structure of these systems networks, which can be applied to genetic association studies. Computational representations are now available for a number of biochemical pathways,[78] modeling detailed (mostly intracellular) interactions between proteins, genes, and small molecules. One relevant example is the nicotine metabolism pathway now available in the HumanCyc[79] and PANTHER Pathways[80] databases. Higher-level representations are also available that model the relationships between various "constructs" (concepts) in a field, such as nicotine dependence, withdrawal, and smoking relapse.[81] These data sources can be used to define a network structure that relates genetic and environmental variables to phenotypes (and endophenotypes) in a detailed manner, as illustrated in figure 12.2B. This network differs from the network in figure 12.2A in two main aspects. First, the actual number of edges (connections) can be much smaller than in the "hypothesis free" network, which reduces the "search space" of likely models for genotype-phenotype effects. Second, there are intermediate nodes between

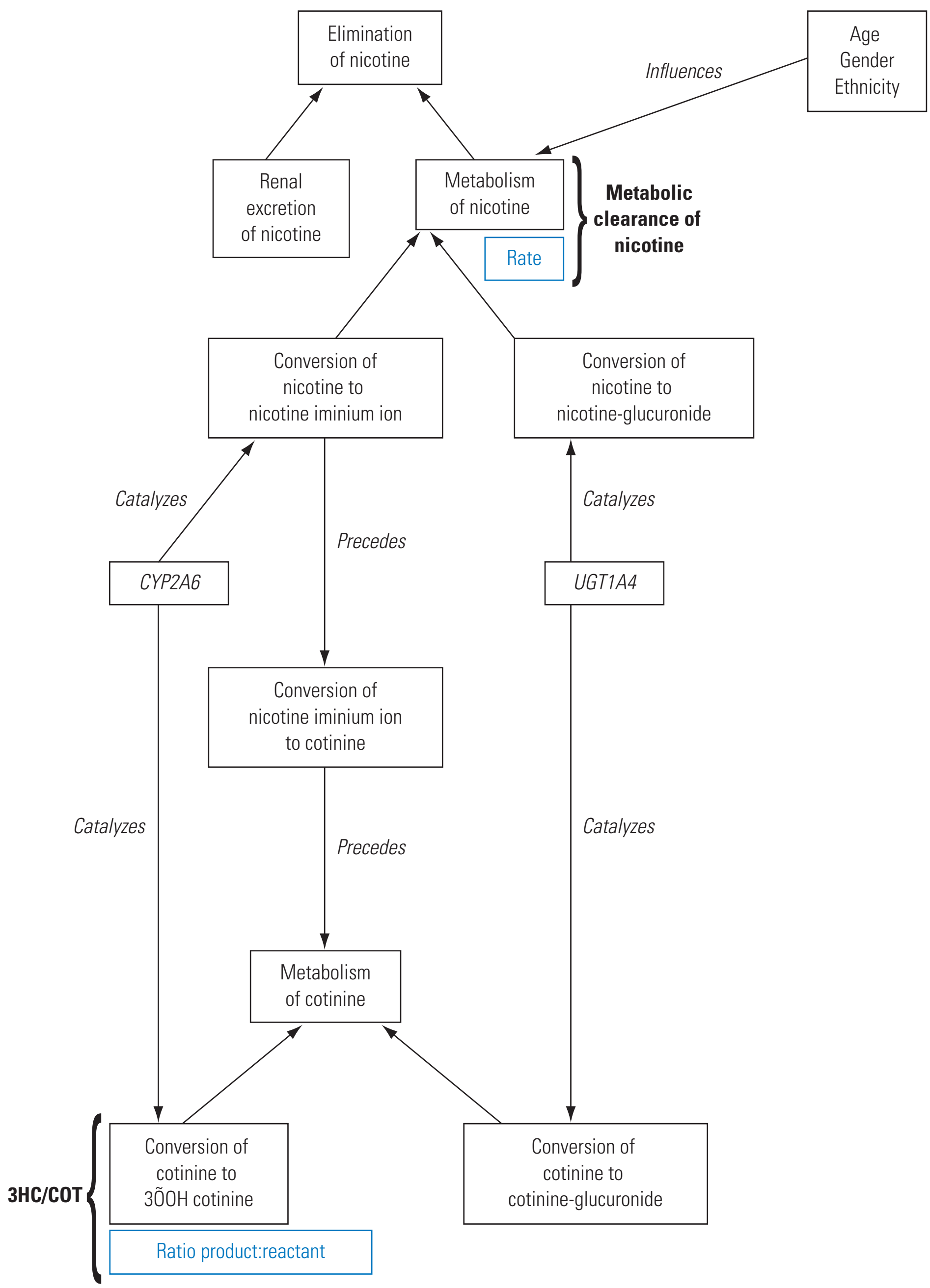
Figure 12.2 Examples of Networks



Note. Shown is a generalized network of genes and all possible relations within the nicotine metabolism pathway. 3HC/COT = *trans* 3′-hydroxycotinine to cotinine ratio.

Figure 12.2 Examples of Networks *(continued)*

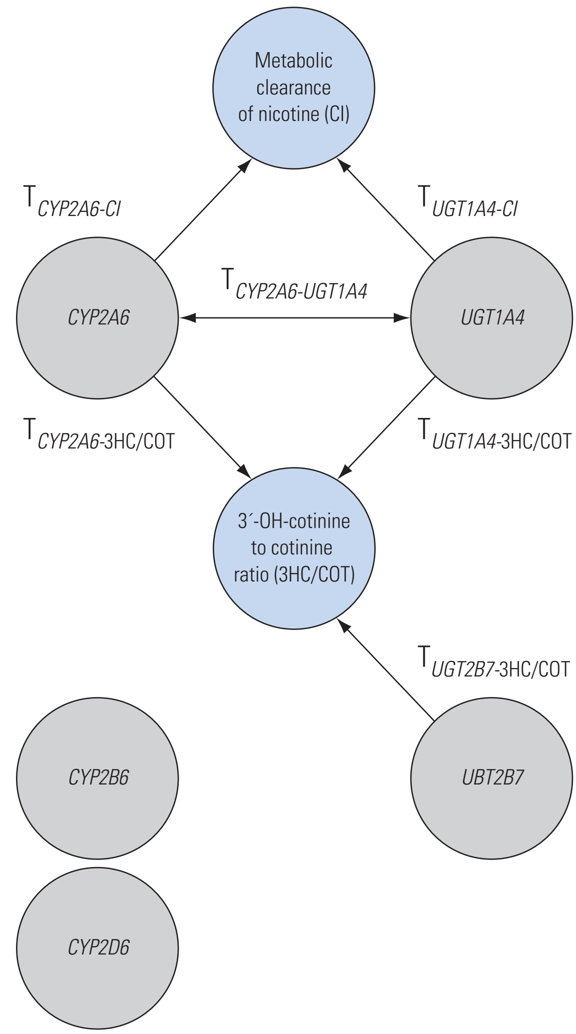
B.



Note. Shown is a simplified Nicotine Pharmacokinetics Ontology. 3HC/COT = *trans* 3′-hydroxycotinine to cotinine ratio.

Figure 12.2 Examples of Networks *(continued)*

C.



Note. Shown is the network characterization of the effect of each gene (via the test statistic T) and its impact on the outcome of interest (either nicotine clearance or 3′-OH-cotinine to cotinine ratios) and the relations to each other as specified in the ontology.

genotype and phenotype that can serve as "endophenotypes," which can be assayed to model (and validate) mechanistic relationships between genotype and phenotype. For the purposes of this paper, the network structure can be used to estimate prior probabilities of effects and interactions that are different for different genetic and environmental variables, as well as for different interaction terms.

Using Ontologies to Represent Prior Knowledge About Relationships Between Variables

In genetic association studies, there is often prior knowledge of relationships that is applied to hypothesis testing. The most commonly used relationships are physical or genetic distance-based relationships between individual single-position genotypes, such as LD, or, for functional relationships, presence in the same gene or in a list of "candidate genes." Candidate genes are generally selected using prior knowledge, such as previously reported associations with the same or similar phenotypes or known or hypothesized biological relationships. Here, it is suggested that ontologies provide a useful formalization of these relationships, enabling the incorporation of multiple types of prior knowledge into computational analyses of genetic association data.

The goal of genetic association studies is to uncover statistical correlations between genetic (germline) variation and phenotypic variation. To be useful for genetic association studies, an ontology must represent concepts (or "entities") in the domains of both genotypes and phenotypes and the relations between these concepts.

An ontology is a formal structuring of knowledge.[82] For the purposes here, an ontology is a *formal model* of a domain of knowledge, and consists of *entities* and *relations* between entities. An entity is simply a class, or category, of things that one wishes to model. An entity can be either a *continuant* (an object existing at a particular point in time) or an *occurrent* (an event or process occurring over time). Relations can be of many types, depending on the knowledge domain being represented, but two of the most common are the "is_a" relation, which specifies one class as a subclass of another (for example, human is_a mammal), and the "part_of" relation (e.g., finger part_of hand). Ontologies have their origins in Aristotelian philosophy, but computer science has driven a renaissance in ontology development and use—namely, by the problem of representing computational knowledge in the artificial intelligence field and the Semantic Web.

Why Build an Ontology?

Formalizing a particular knowledge domain can have two main impacts on a scientific research field. First, it can help to *clarify* and *communicate* the (often) implicit models used by scientists to formulate and test hypotheses. It clarifies the models by making them formal and explicit. Trying to formalize an implicit model can often be a useful exercise in itself, but for knowledge domains that are too large or complex for a single scientist to be an expert in all relevant subdomains, it is critical. A structured representation can help to clearly communicate a model to other researchers and allow iterative community development and revision of the model.

Second, an ontology can make expert domain knowledge accessible to *computation*. A computer may not yet "understand" the knowledge (in the human sense, whatever that means), but it can take advantage of the relations between entities in computational models and numerous useful algorithms, such as those aiding humans to find relevant information on the Web. The focus in this paper is on one such application: ontologies can facilitate the building of computational models for the testing of genotype-phenotype associations.

Clearly, then, ontology development will have the greatest impact on a given scientific area of inquiry when the field is sufficiently complex to be beyond the expertise of (most) single researchers. Such fields are interdisciplinary almost by definition, including or depending on many subfields of specialized expertise. As noted by Karp,[83] a scientific theory can be structured "within a formal ontology so that it is available for computational analysis." The resulting *computational symbolic theory* enables "analysis and understanding for theories that would otherwise be too large and complex for scientists to reason with effectively."

How To Build an Ontology Relating Genotypes and Phenotypes

Ontologies have been developed for a number of biomedical domains, including anatomy, gene function, biochemical pathways and mutant and strain phenotypes in experimental model organisms such as mouse, zebrafish, fruit fly, and yeast. The experiences of these groups, as well as groups from other domains, have led to a number of proposed best practices for ontology development.[84–88] Here, the focus is on building an ontology that relates genotypes and phenotypes. A well-known ontology development process[89] has been adopted for this purpose.

Step 1: Determine the domain and scope. A genotype-phenotype ontology must include concepts from the molecular level (such as gene and genetic variation) to the phenotypes (such as a disease), including any intermediate-level concepts that may bridge the genotype-phenotype gap (such as biochemical pathways, or particular cell types or organs). Obviously, the concepts will be specific to the phenotype(s) of interest. This is a good point in the process to define "competency questions";[90] these are questions that the ontology, once completed, should be able to address. For a genotype-phenotype ontology, the competency questions should cover such areas as

- What are the prevailing models for disease etiology?
- What are the relevant phenotypes/endophenotypes?
- What biological processes are thought, or hypothesized, to be involved?
- What is known at the molecular level about these biochemical pathways and underlying genes?
- Are there any clues from previous association studies, or from linkage or twin studies?
- What are possible confounding/environmental factors?

Step 2: Consider using existing ontologies. As mentioned above, a number of ontologies exist already in the biomedical domain. One of the best sources for existing ontologies is the Open Biomedical Ontologies (OBO) project.[91] Existing OBO ontologies cover many relevant domains such as anatomy (Foundational Model of Anatomy[92]), biological processes (Gene Ontology[93]), molecular "events" such as biochemical reactions (Event Ontology [EVO][94]), phenotype-directed qualities (Phenotype and Trait Ontology [PATO]), sequence types and features, such as genes and genetic variation (Sequence Ontology[95]), and human disease (Disease Ontology [DO]). OBO ontologies are completely open, and most have ongoing active discussion groups and a process for community maintenance and expansion of the ontology. Of the ontologies mentioned above, all but EVO and DO are also part of the OBO Foundry project, which ensures adherence to strict principles of ontology development.[91]

Step 3: Enumerate important terms. This step involves simply listing terms that are important in the domain of interest. At this point, it is not necessary to decide whether these terms will become entities (a class, or category desirable to model) or qualities (inherent "attributes") of entities. These terms will help to refine the scope of the ontology and to provide the basis for formalizing ontology.

Step 4: Define entities (classes) and relationships between entities. At this stage, one begins to define entities and relationships between them. When possible, terms from existing ontologies should be used. When a new entity is introduced, it is critical that a definition also be provided, to ensure that the term can be interpreted correctly (preferably even by a nonexpert). Most of the necessary relationship types already exist in the OBO Relation Ontology, although one additional relationship, *influences,* was found to be useful for describing putatively causal relations between entities that are critical for a model of the existing domain knowledge. For example, in the Nicotine Pharmacokinetics Ontology given here (NPKO, figure 12.2B), age *influences* metabolism_of_nicotine.

A number of software packages are available for simplifying the task of constructing ontologies. The added benefit of using one of these packages is that at the end of the process, the ontology is stored in a standard ontology format, such as the OBO format. As a result, the ontology can be imported into a number of software tools, such as those developed for the Ontology Web Language, or OWL,[84,96] for analyzing the ontology for consistency and for computational reasoning over the ontology. Among the most popular packages for developing biomedical ontologies are OBO-Edit[97] and Protégé.[98]

For biochemical pathways, the BioPAX Ontology[99] is beginning to enter widespread use. Well-known pathway resources such as BioCyc,[100] Kyoto Encyclopedia of Genes and Genomes,[101] Reactome,[102] and PANTHER[103] have made a relatively large number of pathways available in BioPAX format[99] and SBML (Systems Biology Markup Language).[104] SBML has the advantage of being able to specify quantitative data such as reaction rate constants, but BioPAX has greater expressive capability for genomic and protein sequence data that is critical for treating genetic variation data. If a relevant pathway does not yet exist in sufficient detail in one of these resources, PANTHER Pathways has a community curation Web site where domain experts can take advantage of the CellDesigner[105] program's interface to draw a pathway and store a formal ontology representation directly from the drawing.

Step 5: Define qualities important for representing phenotypes.

Once the entities are defined, *qualities* can be enumerated for each of the entities. The emerging standard for phenotypes is the PATO syntax. In this "bipartite" entity:quality definition, a phenotype (e.g., metabolic clearance of nicotine) is expressed as a combination of an entity (e.g., metabolism of nicotine) and a quality inherent in the entity (e.g., rate). Phenotypes can be quantitative or qualitative. For example, a particular chemical reaction type (entity) might have a rate (quality), which would then be specified by a particular quantitative measurement (value).

Most ontology development projects begin with the formation of a small working group that brings together expertise in the relevant knowledge domain, with expertise in ontology construction. In the biomedical field, the National Center for Biomedical Ontology (NCBO) has been established; one of its primary missions is to provide the ontology construction expertise to facilitate development of new ontologies for biomedical applications.[106] The NCBO is an excellent resource for expert guidance and software tools for this purpose.

The product of the initial working group is a draft ontology. If appropriate, this draft ontology can provide a framework and starting point for a larger, community-driven project to expand and refine the ontology. At this point, the ontology enters a completely new phase of development. Community projects such as this require an infrastructure for managing discussions to come to a resolution on proposed changes and then rapidly incorporate accepted changes to the ontology. The OBO project provides a platform for facilitating community ontology projects, leveraging resources originally designed to support Open Source software development projects, such as the SourceForge Web site.[107]

Example: Nicotine Metabolism

Data

As an example, data are used from a study involving the volunteer-based Northern California Twin Registry.[1,2] This study of the heritability of nicotine metabolism included 278 individuals between the ages of 18 and 65 years. Individuals were excluded for the following: greater than 30% above normal weight range; pregnancy; use of known drug metabolism-altering medications (e.g., barbiturates, phenytoin, rifampin [or INN, rifampicin]); uncontrolled hypertension or diabetes; heart, lung and cardiovascular disease; cancer; liver and kidney diseases; substance abuse or dependence; positive human immunodeficiency virus status; history or evidence of hepatitis B or C; and discomfort with venipuncture procedures. Both nonsmokers and smokers were recruited. Further details regarding the study description can be found elsewhere.[1,2] Quantitative data were obtained to measure the impact of genetic variants on the disposition kinetics and metabolism of nicotine after systemic administration. As such, participants of the twin study were administered intravenous deuterium-labeled nicotine and cotinine (the major proximate metabolite of nicotine) and blood samples were obtained for genotyping. From blood concentrations obtained at intervals over 72 hours and urinary excretion data, pharmacokinetic parameters were estimated using model-independent methods.[108] Here, attention is confined to two outcomes of interest: the total clearance of nicotine, and *trans* 3′-hydroxycotinine to cotinine ratio (3HC/COT). *Trans* 3′-hydroxycotinine is the major metabolite of cotinine, and its formation is catalyzed almost or entirely exclusively by *CYP2A6,* the enzyme that is primarily responsible for the metabolism of nicotine. The 3HC/COT ratio has been used as a marker of *CYP2A6* activity and of the clearance rate of nicotine.[109] Because this data set has a limited number of smokers, and previous analyses have demonstrated that inference for pharmacokinetics of nicotine remained largely unchanged after controlling for smoking status, smoking status is not included in the present analysis for simplicity. The analysis is limited to only Caucasians ($N = 211$), and age is included as the only covariate for demonstration purposes. Genotypes available for analysis include "wild-type" *CYP2A6*1* and its most common variants: *CYP2A6*2, CYP2A6*4, CYP2A6*7, CYP2A6*8, CYP2A6*9, CYP2A6*10, CYP2A6*12;* four SNPs within *CYP2B6;* a single SNP within *CYP2D6;* seven SNPs in *UGT1A4;* and four SNPs in *UGT2B7.*

Analysis

To begin, a univariate analysis was performed by comparing the means for the kinetic parameters by each variant by using a mixed linear model for the first-stage likelihood, $f\left(Y \mid \mathbf{X}, \underline{\beta}, \underline{\gamma}\right)$, in which a random effect is included for twins to control for nonindependence. For *CYP2A6,* a previously reported analysis was followed,[108] and three categories were created on the basis of the impact of individual genotypes on nicotine clearance, fractional clearance, cotinine clearance and the 3HC/COT ratio: (A) **1/*1;* (B) **1/*9* or **1/*12;* and (C) any of the following variants: **1/*2, *1/*4, *9/*12, *9/*4, *9/*9* (*CYP2A6*7, *8, *10* were not found in this data set). Thus, the linear model has two dummy variables for groups (B) and (C), reflecting the difference in means relative to the referent group (A). For the remaining SNPs in the other genes, an additive coding representing the number of variant alleles was used. For the three SNPs with individuals with missing genotypes, the expected coding was substituted as a function of allele frequency. While this may result in an underestimated variance, one should not expect an appreciable difference in that the number of individuals with missing values is small ($N = 1$, 6, or 7, respectively). Age is included as a continuous covariate in every model.

For the hierarchical stochastic search, the first step was to outline a full model in which there are 18 main effects for gene polymorphisms (two dummy variables for *CYP2A6* and 16 SNPs across the other four genes), one main effect for age, and 170 pairwise interaction terms that include within and across gene interactions and gene-by-age interactions. For interpretability, the two dummy variables for *CYP2A6* are always included in the model together. Because the SNPs within a single gene were in relatively low LD, only the conventional interaction term (i.e., a deviation form additivity) was modeled and a phase term as described in equation (5) was not created. For the stochastic search, a hierarchical constraint on interaction terms was included, allowing interactions in the model only if their parental main effect terms are included. For this example, interaction terms were included to illustrate the feasibility and computational challenge of searching over a substantial model space. However, in this particular application, one should not expect to be able to detect interaction effects because of the limited sample size. Under favorable assumptions for two genes interacting (i.e., common allele frequencies and a large effect size—comparable to that observed in Benowitz and colleagues[108] for the main effect of *CYP2A6*)—the power to detect an interaction with this sample is about 10%–20%.

Ontology and Incorporation into the Hierarchical Stochastic Search Model

An Example Ontology Linking Genotypes and Phenotypes for Nicotine Pharmacokinetics

As part of the Pharmacogenetics of Nicotine Addiction and Treatment project funded by the National Institute on Drug Abuse, the authors of the chapter are developing a draft ontology in the areas of nicotine pharmacokinetics, dependence, and treatment outcomes. Figure 12.2B shows part of the initial draft of the NPKO relevant to the outcome phenotypes addressed in this paper. The ontology has several notable properties. First, it is hierarchical (more properly, the structure is a directed acyclic graph, or DAG, meaning that a child class can have more than one parent). Second, it spans the range from genotype to phenotype, representing high-level phenotypes, intermediate-level "endophenotypes" down to molecules and genotypes. Third, phenotypes are represented using the emerging PATO standard,[110] shown as two adjacent ontology terms, an *entity* (black typeface) and a *quality* (blue typeface) in figure 12.2B.

Using the Nicotine Pharmacokinetics Ontology to Derive Priors

A discussion follows on how the information encoded into the ontology can help to define priors in the context of the Bayesian model selection process outlined previously.

What does figure 12.2B reveal in terms of prior information regarding the influence of genes on the phenotypes? In other words, how might the different effect estimates as summarized in the test statistics be related to each other? The first phenotype, 3HC/COT, is the ratio of the concentrations of 3HC and cotinine, and therefore variation in any genes connected in the network (figure 12.2B) to either 3HC or cotinine, or both, could have an effect on this ratio. *CYP2A6* catalyzes the conversion of 3HC to cotinine, which would clearly be expected to have the primary effect on the 3HC/COT ratio. However, since *UGT1A4* activity depletes cotinine by conversion to cotinine-glucuronide and *UGT2B7* activity depletes 3HC by conversion to 3HC glucuronide, variation in both *UGT1A4* and *2B7* could also affect the 3HC/COT ratio.

The second phenotype, metabolic clearance of nicotine, relates to the rate at which nicotine is converted to other compounds. In the simplified NPKO (figure 12.2B), there are two pathways for nicotine metabolic clearance: nicotine can be converted into either cotinine or nicotine glucuronide, reactions catalyzed by *CYP2A6* and *UGT1A4,* respectively. The ontology, therefore, specifies that variation in both *CYP2A6* and *UGT1A4* would be expected to affect nicotine metabolic clearance. One can use further prior information—namely, in most individuals, more nicotine was found to be metabolized through the cotinine pathway than the nicotine-glucuronide pathway, by a factor of about 15,[111] to specify the prior belief that *CYP2A6* variation will have a larger effect on nicotine metabolic clearance than does *UGT1A4.* The relative rates of these reactions are stored in the ontology in the following form:

conversion_of_nicotine_to_nicotine_iminium_ion:relative_ _rate

Compar conversion_of_nicotine_to_nicotine-glucuronide

M 15,

where Compar denotes "in comparison to" and M denotes "measurement," using the PATO standard terms.

The relations between genes and phenotypes, represented in the ontology (figure 12.2C), therefore provide a list of nonzero priors for the effects of variation in each gene on each of the phenotypes. They also provide expected relative contributions to the phenotype; namely, *CYP2A6* is expected to have the primary effect on both 3HC/COT and nicotine metabolic clearance. For simplicity, the expected effect of *CYP2A6* on the 3HC/COT ratio was set to be four times as large as the expected effect of either *UGT2B7* or *UGT1A4.* The ontology can also provide prior effect estimates for gene-gene interactions. *CYP2A6* and *UGT1A4* are both involved in the two phenotypes, nicotine clearance and 3HC/COT, so the gene-gene interaction term is expected to be nonzero for these two genes in both phenotypes.

Finally, relatively little is known regarding the specific polymorphisms within each gene, so a single prior value applicable to all SNPs within a gene is assigned. Taken together, the ontology yields the following matrix of priors:

Gene	Metabolic clearance of nicotine	3HC/COT ratios
CYP2A6	4	4
CYP2B6	0.5	0.5
CYP2D6	0	0
UGT1A4	1	1
UGT2B7	0	1
CYP2A6-UGT1A4	1	1
All other interactions	0	0

Incorporating Priors into Statistical Analysis

In addition to the above prior covariates for each respective analysis, an intercept term and a dummy prior covariate are included for main effects versus interaction effects. The same prior covariates are used for both the means and probability portions of the mixed model. Furthermore, in the means model, the intercept of the noncentrality parameter e_v is constrained to be equal to the expectation of a chi distribution under the null of no associated terms in the regression model for identifiability. This causes the interpretation of the remaining effects of the prior covariates on the magnitude—that is, the $\boldsymbol{\mu}$s—to reflect a deviation from the null expectation. In the probability portion of the model, $\pi_0 = 1$ and $\pi_1 < 0$ are constrained, corresponding to the effects of all the terms and the main effects on the probability of inclusion. These constraints limit the inclusion of main effects via π_1 and thus guide the stochastic search to more parsimonious models in terms of the number of main effects included in the model. This is important in that the relatively small sample size in this example ($N = 211$) prohibits the fitting of models with too many main effects. However, once a set of main effects is included in a model, one wants to encourage the exploration of models with interactions. Thus, by setting $\pi_0 = 1$, the expectation of the inclusion of an interaction conditional on the inclusion of the parental main effects is relatively high (21).

Sensitivity to Prior Specification

To compare and investigate the sensitivity of inference to the prior covariate specification, two alternative specifications are used. First, the above prior covariate matrix is altered by the assumption that *CYP2A6* has the same impact as *UGT1A4*. This is accomplished by replacing the "4" with a "1" in the previously described prior covariate matrix. Second, in assuming that the prior knowledge is limited, a prior covariate design matrix is used with five dummy variables indicating the gene in which a specific polymorphism is found. Here, it is assumed that all the polymorphisms within a gene are exchangeable or share a common mean with a different mean for each gene. This allows sharing among polymorphisms within a gene, but not across genes.

Results

Univariately, polymorphisms for groups (B) and (C) for *CYP2A6* were significantly associated with measured nicotine clearance levels as seen in table 12.1 ($t_B = 2.15$, p-value = 0.03; $t_C = 3.86$, and p-value = 0.0002, respectively). In addition, SNP 4 within *UGT1A4* had a statistically significant result ($t_{SNP4} = 2.19$, p-value = 0.03). Because of the small sample size, a model could not be fitted in which all possible polymorphisms were included as represented in equation (2), thus limiting any further exploration of full joint models with interactions without some type of model selection procedure.

The hierarchical stochastic search model was implemented by using the statistical software R.[112] Posterior inference was based on 50,000 samples from a single chain after discarding the first 10,000 samples (i.e., burn-in) to ensure that the final inference is independent of the starting values.[113] Visual inspection of time series and sensitivity to inference over time was used to check for convergence and model performance. The burn-in period was selected because it was found that the constraints in both the means and the probability portions of the model allowed for a nonzero probability of including any given main effect in the model. This results in sufficient mixing within the model space and a very limited dependence on the starting model. For example, under the null of no association between any

$$\Pr(\gamma_{m\partial} = 1 \mid \gamma_m = 1, \gamma_\ell = 1) = \exp(0) \big/ (1 + \exp(0)) = 0.5 \qquad (21)$$

Table 12.1 Results for Nicotine Clearance

		Univariate		Hierarchical model			
Gene	**Variant**	T[1]	*p*-value	$\tilde{E}_v$[2]	$\tilde{P}_v$[3]	$BF(\tilde{P}_v)$[4]	$BF(\gamma_v = 1)$[5]
CYP2A6	Group A (*1/*1)	—	—	—	—	—	—
	Group B (*1/*9 or *1/*12)	2.15	0.03	3.53	0.09	3.16	17.38
	Group C (*1/*2; *1/*4; *9/*12; *9/*4; *9/*9)	3.86	0.0002	3.53	0.94	534.37	17.38
CYP2B6	SNP 1	1.22	0.23	2.14	0.02	0.62	0.82
	SNP 2	1.82	0.07	2.27	0.04	1.33	0.82
	SNP 3	1.56	0.12	1.78	0.04	1.36	0.74
	SNP 4	1.69	0.09	2.32	0.03	0.90	0.82
CYP2D6	SNP 1	0.21	0.83	1.51	0.02	0.49	0.70
UGT1A4	SNP 1	0.38	0.70	2.19	0.01	0.40	0.84
	SNP 2	0.02	0.98	2.18	0.01	0.40	0.84
	SNP 3	0.07	0.94	1.64	0.02	0.56	0.81
	SNP 4	2.19	0.031	2.01	0.08	3.90	0.76
	SNP 5	0.55	0.58	1.69	0.02	0.52	0.77
	SNP 6	1.57	0.12	2.43	0.02	0.75	0.81
	SNP 7	0.81	0.42	2.16	0.02	0.65	0.84
UGT2B7	SNP 1	1.22	0.23	1.58	0.03	0.84	0.73
	SNP 2	0.12	0.90	1.83	0.02	0.42	0.76
	SNP 3	0.50	0.62	1.53	0.02	0.54	0.67
	SNP 4	0.05	0.96	2.23	0.02	0.58	0.73

Note. Results were obtained by using a conventional univariate regression analysis and from the hierarchical stochastic search by using informative prior covariates derived from the ontology.

[1]The absolute value of the χ test statistic obtained from the Wald-type test from a univariate regression model.

[2]Posterior expectation of the χ test statistic.

[3]Posterior expectation of the probability that the association is true.

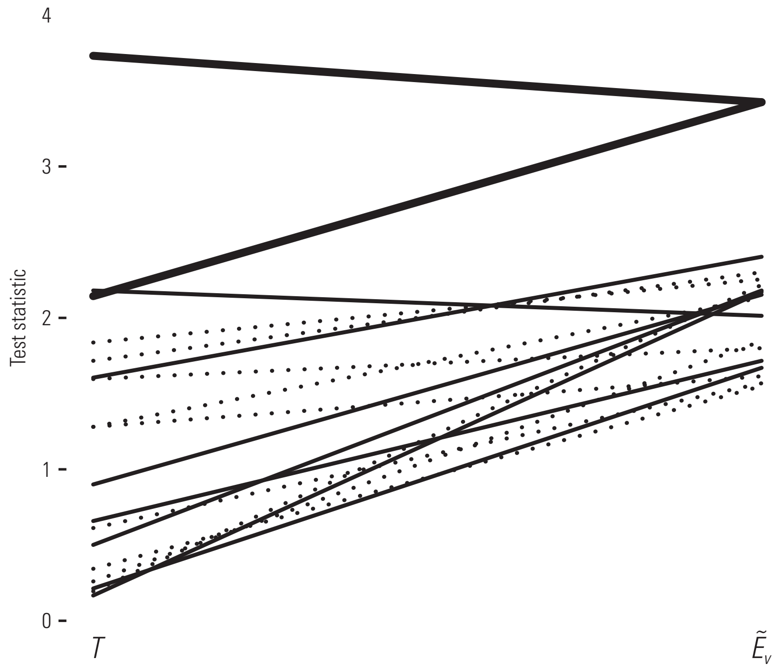
[4]Bayes factor for the probability that the association is true.

[5]Bayes factor for the inclusion of the corresponding term in the regression model.

polymorphisms and nicotine, the average probability of including any term was 3%. This encourages sampling the model space, but since the probability of including a term is nonzero under the null, it also highlights the need to compare posterior estimates of the probability of a true association and the probability of including a given term conditional on the data to those under the null via Bayes factors to obtain valid inference. To guarantee sufficient mixing within the local model space, a random walk over 500 iterations was incorporated in which an additional main effect to the model under evaluation was included.

The focus initially is on the posterior estimates of the magnitude of the association $\tilde{E}_v$ for nicotine clearance. Recall that the ontology specified that the two groups of polymorphisms in *CYP2A6* would have twice the effect of polymorphisms in *UGT1A4* and that there would be no effect for all other variants in other genes.

Figure 12.3 Shrinkage of Test Statistics



Note. Shown is the shrinkage of the univariate test statistics *T* by the hierarchical stochastic search to yield the posterior estimates of effect size $\tilde{E}_v$. For each term, the test statistic obtained from the univariate analysis is paired with the posterior estimate, demonstrating shrinkage to a conditional mean specified by the prior covariate structure.

This structure is reflected in the posterior estimates for $\tilde{E}_v$ summarized in table 12.1. The two groups of polymorphisms in *CYP2A6* have similar posterior estimates of $\tilde{E}_{CYP2A6,B} = 3.53$ and $\tilde{E}_{CYP2A6,C} = 3.53$. Similarly, the estimates for the posterior magnitude of the test statistics for SNPs within *UGT1A4* are shifted toward each other, albeit at a lower magnitude. The combined effect of the prior covariates is more clearly seen in figure 12.3, which demonstrates the shrinkage of the original test statistic to the posterior estimates. First, by grouping all main effects via the second covariate in $\mathbf{Z}_\mu$, all the posterior estimates are shrunk upward toward a group effect. Furthermore, within this upward shrinkage, the prior covariate based on the ontology allows further borrowing of effects within *CYP2A6* (the bold solid lines) to be four times the magnitude of that of the SNPs within *UGT1A4* (the thin solid line). All other polymorphisms (dashed lines) have upward shrinkage based solely on the grouping of main effects.

In focusing on the posterior estimates of the probability of a true association $\tilde{P}_v$, it can be seen that *CYP2A6* group C has a much larger probability of being true ($\tilde{P}_{CYP2A6,C} = 0.94$) in comparison with group B ($\tilde{P}_{CYP2A6,B} = 0.09$) despite their similar posterior estimates for $\tilde{E}_v$. This is due to the contribution of the data for each polymorphism group reflected through their corresponding first-stage test statistic of $T = 3.86$ and $T = 2.15$, respectively. Furthermore, because the posterior probability is not zero under the null, inference into the significance of this estimate should be made via the Bayes factor. Here, very strong evidence can be seen for a true association for group C in

CYP2A6 ($BF(\tilde{P}_{CYP2A6,C}) = 534.37$) as well as positive evidence for an association of group B in *CYP2A6* ($BF(\tilde{P}_{CYP2A6,C}) = 3.16$). In addition, an indication can be seen for positive evidence of an association for SNP 4 in *UGT1A4* with $BF(\tilde{P}_{UGT1A4,4}) = 3.90$. Although these conclusions are qualitatively similar to conclusions based on the results obtained from the univariate analyses, there are some notable differences. For example, the test statistics obtained for group B in *CYP2A6* and for SNP 4 in *UGT1A4* are similar, suggesting comparable evidence for an association. However, they have very different posterior estimates for $\tilde{E}_v$ with $\tilde{E}_{CYP2A6,B} = 3.53$ and $\tilde{E}_{UGT1A4,4} = 2.01$, reflecting the borrowing of information via the prior structure; mainly, the test statistic for group B is shrunk upwards toward group C within *CYP2A6*. When one accounts for the influence of the prior structure and focuses on the Bayes factors for a true association, there is slightly more evidence for SNP 4 in *UGT1A4*: $BF(\tilde{P}_{UGT1A4,6}) = 3.90$ versus $BF(\tilde{P}_{CYP2A6,B}) = 3.16$. The evidence for group B is tempered because of the strong prior for the influence of *CYP2A6*—that is, it would have a fourfold increase in effect. In contrast, it was believed that *UGT1A4* would have a much smaller effect, and thus, the impact of the data relative to the prior is greater.

The hierarchical stochastic search model did not find any evidence for interactions between polymorphisms or between the polymorphisms with age. Most likely, this is mainly a reflection of the limitations for obtaining statistical significance for interactions with such a small sample size ($N = 211$) of a narrow age range. However, one of the major goals of incorporating prior knowledge was to have an efficient stochastic search across the model space. In this regard, guiding the stochastic search via the proposal distribution as a function of the probability that an association is true results in a very high acceptance rate during the MCMC iterations (across the various analyses, on average 90% of the proposed models are accepted). At first glance, this high acceptance rate may indicate poor mixing in the MCMC chain, leaving one unable to move around in the model space. To some extent, the exploration of the entire space is limited, but sampling of models believed to be more biologically plausible is actively encouraged. Specifically, the prior structure given from the ontology indicates the desirability of investigating interactions between *CYP2A6* and *UGT1A4*. As evidence of a guided search, it was found that, conditional on the inclusion of the two polymorphism groups within *CYP2A6*, interactions with SNPs within *UGT1A4* are included in 3% of the models searched. In contrast, in performing a stochastic variable selection and substituting a binomial proposal distribution that is not dependent on a prior structure but has probabilities reflective of the hierarchical model under the null, it is found that interactions between *CYP2A6* and *UGT1A4* are included in less than 0.1% of the models searched.

To gauge the sensitivity of the results to prior specification, two additional analyses were run using different prior covariates. To mimic the influence of incorrect priors and assuming a lack of knowledge for *CYP2A6*, a “1” was substituted in place of the “4” for *CYP2A6* in the previous prior covariate matrix. This had little impact on the final inference in regard to the posterior estimates corresponding to *CYP2A6*, further indicating that the data are driving the results for *CYP2A6*. However, under this prior structure, estimates for polymorphisms within *UGT1A4* were slightly attenuated because they were no longer shrunk upward toward the *CYP2A6* estimates, for SNP 4 $\tilde{E}_{UGT1A4,4} = 1.73$. Despite the change in estimates, the Bayes factor for the posterior probability of a true association still indicated some positive evidence for SNP 4, $BF(\tilde{P}_{UGT1A4,4}) = 3.81$. With the gene-specific prior covariate matrix that includes a set of dummy variables indicating which gene an SNP is in, qualitatively similar results are found. Of course, inference in terms of both posterior estimate and the magnitude of the Bayes factors varies, reflecting differences in the borrowing of information across polymorphisms as specified by the prior structure (results not shown).

Results for the analysis in regard to 3HC/COT are presented in table 12.2. As before, similar patterns are apparent in posterior estimates with the most notable evidence provided for the two groups of polymorphisms in *CYP2A6*. Of note are the estimates for SNPs in *UGT2B7*. For 3HC/COT, a prior covariate matrix was specified that placed a slight emphasis on *UGT2B7* in conjunction with *CYP2A6*. For the posterior estimation for the magnitude $\tilde{E}_v$, the four SNPs in *UGT2B7* are shrunk upward, reflecting a borrowing of information from the larger *CYP2A6* estimates.

Table 12.2 Results for 3HC/COT Ratios

		Univariate		Hierarchical model			
Gene	**Variant**	T[1]	***p*-value**	$\tilde{E}_v$[2]	$\tilde{P}_v$[3]	$BF(\tilde{P}_v)$[4]	$BF(\gamma_v = 1)$[5]
CYP2A6	Group A (*1/*1)	—	—	—	—	—	—
	Group B (*1/*9 or *1/*12)	-2.53	0.01	4.39	0.09	2.83	13.87
	Group C (*1/*2; *1/*4; *9/*12; *9/*4; *9/*9)	-4.9	4.0E-06	4.39	0.92	291.79	13.87
CYP2B6	SNP 1	-1.47	0.15	1.17	0.06	1.28	1.10
	SNP 2	2.39	0.02	1.35	0.14	3.47	1.02
	SNP 3	1.31	0.19	1.16	0.04	1.02	0.64
	SNP 4	2.18	0.03	1.22	0.09	2.25	1.05
CYP2D6	SNP 1	-0.51	0.61	1.32	0.03	0.58	0.86
UGT1A4	SNP 1	0.13	0.90	0.95	0.03	0.71	1.03
	SNP 2	-0.68	0.50	1.00	0.03	0.77	1.03
	SNP 3	0.93	0.35	0.98	0.04	1.03	0.97
	SNP 4	1.56	0.12	1.02	0.05	1.19	0.86
	SNP 5	0.89	0.38	0.95	0.03	0.80	1.00
	SNP 6	1.55	0.12	0.97	0.04	1.20	1.06
	SNP 7	-0.34	0.74	0.94	0.03	0.71	1.00
UGT2B7	SNP 1	-0.94	0.35	1.03	0.03	0.87	1.05
	SNP 2	0.16	0.87	0.86	0.03	0.80	0.97
	SNP 3	0.6	0.55	0.95	0.04	1.08	1.05
	SNP 4	-2.49	0.01	1.36	0.17	4.90	1.16

Note. Results were obtained by using a conventional univariate regression analysis and from the hierarchical stochastic search by using informative prior covariates derived from the ontology. 3HC/COT = *trans* 3′-hydroxycotinine to cotinine ratio.

[1]The absolute value of the χ test statistic obtained from the Wald-type test from a univariate regression model.

[2]Posterior expectation of the χ test statistic.

[3]Posterior expectation of the probability that the association is true.

[4]Bayes factor for the probability that the association is true.

[5]Bayes factor for the inclusion of the corresponding term in the regression model.

Discussion

Statistical modeling does have limits, especially when evaluating multiple exposures and genes with a limited sample size. In light of these limitations, conventional univariate analyses can be appealing in their ease of implementation and straightforward interpretation. However, building upon the knowledge that guided the initial selection of the SNPs and genes for investigating, most researchers feel compelled to go beyond the independent treatment of each gene and attempt to model more complex joint action and interactions. Often, this includes ad hoc criteria for model building on the basis of prior biological knowledge with the analyst balancing the complexity of each model investigated with real world limitations of the data, such as multicolinearity, sparse data bias, and instability. Rarely do final models accurately reflect the statistical costs in terms of multiple comparisons or the uncertainty in arriving at a given "best" model. As an alternative, the analyst may opt to use strictly data-driven approaches and search for significant interactions by using statistical criteria. Within this context, the method presented here represents the use of a hierarchical model together with a means of using prior knowledge to guide statistical model selection by means of an ontology.

The idea of placing more emphasis on more biologically relevant SNPs is not new. Several other approaches have been presented. The false positive report probability uses prior information in the form of an investigator's prior belief that an association is true. Likewise, a weighted FDR and Bayesian FDR approach have been presented to incorporate outside information on the a priori impact of a particular SNP. However, these approaches rely on prespecification of the weight or prior for every SNP and interaction term without allowing the data to enhance or attenuate the influence of the prior information. In contrast, the hierarchical modeling approach discussed here relies on prespecification of only *how* it is believed that SNPs and interaction terms are related, but it relies on the data to determine *the degree* or the weight of the various specifications or prior covariates. This has the advantage of giving some flexibility in the prior specification, and correspondingly, final inference and conclusions may be less sensitive to those specifications. Thus, the posterior estimates for the importance of each term and interaction are conditional on the prior knowledge, and within this modeling framework these parameters are naturally interpreted in the context of that knowledge. This avoids having post hoc justification and rectification of conventional results with what is known. As knowledge changes, the analyses can be rerun to gauge how new knowledge combined with the sampled data may alter final conclusions. While sensitivity analysis is a vital part of any comprehensive Bayesian analysis, with subjective priors one does not expect the results to be quantitatively similar across a variety of prior structures. In fact, the goal is just the opposite. One would like to use subjective knowledge as a guide to models that would not have been found otherwise or to enhance posterior estimates that may have been overlooked without shrinkage to other SNPs or genes. But, one must also be careful that the final inference does not solely reflect specific prior beliefs. The use of Bayes factors gauges the evidence for the conclusions conditional on the data and in the context of the priors.

Ultimately, it is a fine line between deterministic weights and informative priors. The authors of this chapter believe that this line is drawn by the quality of the prior information. While much has been done with hierarchical modeling in epidemiological analysis, relatively little research has been done on the quality of the prior covariate specification. Here, an approach is described that attempts to formalize the prior knowledge via an ontology. Ontologies provide a mechanism for investigators to specifically structure their prior knowledge in a usable format. Of course, what is specified in the ontology is not the truth, but only reflective of the available state of knowledge. As such, ontologies can and should be dynamic. In fact, how an ontology changes over time is instructive in indicating areas for advancement and further research.

Ontologies provide a structure for encoding prior knowledge or hypotheses. The existing PATO syntax allows for specifying relationships between concepts and for specifying relative quantities. An example has been given of how both relationships and relative quantities can be used to derive priors in the context of Bayesian model selection, which is, as far as known, a novel application of biological ontologies. The ontology provided a structure for estimating the prior probability that a given gene is involved in a phenotype of interest, as well as the probabilities that different pairs of genes interact with each other.

The part of the NPKO used here is based on extensive evidence from experimental studies, but it would also be possible to encode a more speculative, and even completely untested, hypothesis into an ontology structure to guide model selection. These priors would ensure that the hypothesis will be tested, with high probability, during the model selection step. Of course, whether the hypothesis is accepted will depend on the posterior probabilities after considering the data, and the strength of the evidence as reflected, for example, in the Bayes factors reported here.

In the example given, the ontology structure has been converted into quantitative priors by using expert interpretation. The reasoning followed was simple and could be straightforwardly coded into a computational algorithm. Graph connectivity was used between phenotypes and genes to determine which priors would be nonzero: if a gene was closely connected to the phenotype of interest, the prior was set to be greater than zero. Relative measurements (of reaction rates, in this case) was also used from previous experiments to set the relative values of nonzero priors.

One can expect that one of the most valuable contributions of an ontology for larger studies will be in prioritizing the testing of potential gene-gene and SNP-SNP interactions. The sample data set used was too small to draw any conclusions regarding interactions, but for larger studies that assay a large number of polymorphisms, prioritizing interactions will be critical. Ontologies are one way of estimating a priori probabilities of different interactions. For instance, genes that are closely connected in the ontology relationship network can be hypothesized as being more likely to interact.

Finally, it is straightforward to extend this approach to provide different priors for different individual polymorphisms. For instance, rather than setting the prior expected effects for all polymorphic *CYP2A6* alleles to be the same (relative to the **1/*1* homozygote), functional polymorphism predictions could have been used to provide additional prior information. For instance, allele-specific priors could have been used for *CYP2A6.* The *CYP2A6*9* and *CYP2A6*12* alleles are known to have reduced activity (**9* reduces gene expression through an SNP in the TATA box,[114] while **12* includes exons from the closely related *CYP2A7,* resulting in 10 amino acid substitutions relative to **1* and reduced activity[115]). The *CYP2A6*2* allele[116] has a single amino acid substitution that completely inactivates the enzyme, and, in the *CYP2A6*4* allele, the entire gene is deleted.[117] One could, therefore, have used this prior knowledge (much of which could have been predicted from sequence data alone, e.g., figure 12.1) to specify different priors for the different *CYP2A6* genotypes, with the largest effects expected for individuals having the **4* or **2* alleles. Using functional information about each SNP yields a prior probability that a given SNP will affect gene function. To estimate a prior for the effect of the SNP *on the phenotype of interest,* one could take the product of (1) the conditional prior of the effect of a gene on the phenotype of interest (given an effect on gene function) estimated from the ontology and (2) the prior of the effect of the SNP on gene function.

When including ontological knowledge in statistical analysis, it is desirable to capture potential real world complexities while also addressing the practical limitations of the data—for example, sample size. It is believed that a stochastic variable selection procedure via a hierarchical model offers a potential approach to knowledge-based pathway analyses. Given modeling limitations, one can probabilistically restrict the number of terms included in any specific model via constraints on the conditional probabilities of including a given term. This limits the overall complexity for a regression model evaluated for each iteration of the stochastic search. However, when inference is averaged over all the models, one can begin to describe complex relations between SNPs and genes. In addition, it was demonstrated how prior knowledge can guide the stochastic search efficiently within the model space, yielding more biologically plausible models (in terms of the defined prior covariates). Of course, there is a trade-off of directing the search too narrowly and possibly missing some well-fitting models or of having a broad, nonfocused search in which one may spend most of the stochastic search in an area in which the models are not biologically relevant. Again, this hierarchical framework is a flexible approach that allows multiple sources of information (via the prior covariates) to be included while having the advantage that their actual influence on posterior estimation and the stochastic search does not need to be prespecified but can be estimated from the data.

Details of the specific performance of the statistical model presented here in terms of estimation, sensitivity to prior covariates, ability to identify significant terms, and so on are being pursued in a separate, more statistically oriented paper. While this statistical framework makes use of MCMC methods for the stochastic search across the model space, for computational efficiency maximum likelihood approaches to estimate the first-stage generalized linear model parameters were chosen. Thus, a simplification is made when conditioning on the first-stage maximum likelihood estimates when modeling the second-stage mixture model. Clearly, one can imagine a fully Bayesian analysis in which the uncertainty in the first-stage estimates is propagated into subsequent stages. However, for model selection purposes across such a large model space, it was decided that computational efficiency trumps subtle refinement in estimation. Likewise, the second-stage mixture model uses a maximum likelihood estimation procedure as opposed to a fully Bayesian approach. Again, this decision was made for computational efficiency, and comparisons to the fully Bayesian approach for the mixture model demonstrated suitable performance.[65] With these simplifications, the computations are now on the order of hours as opposed to days with actual times depending on the specific computer. In addition to statistical issues surrounding estimation, there are also issues with how one deals with missing data across all the variables. At the heart of this model selection procedure is a likelihood comparison that requires the likelihoods to be calculated on the same number of individuals. Thus, individuals cannot be removed across models. In the nicotine example, analysis was limited to individuals with complete data or, for the few individuals with missing genotypes, an expected score was imputed. As the number of polymorphisms examined increases, the number of individuals with any missing data will also increase, making this issue a much more serious concern. While the specifics of missing data analysis is beyond the scope of this particular work, the MCMC procedure for model selection provides a flexible framework in which to implement an imputation strategy.

Hierarchical modeling and stochastic variable selection can offer some robustness against multiple comparisons when deciding statistical significance. In 2007, Wakefield[118] formalized the control of false discoveries in genetic epidemiology studies via a prior specification by presenting a Bayesian False Discovery Probability (BFDP). This method is relatively simple to implement and has the advantage of other proposed methods, such as the false positive report probability,[54] by specifying distributions for the null and alternative hypotheses for a given test of association. Furthermore, the BFDP may be calibrated to explicitly incorporate the costs of false discovery versus the costs of nondiscovery. The major limitation of this approach is that it treats each test of association across all polymorphisms as independent. The approach described in this chapter overcomes this limitation by representing a joint distribution over all the test statistics. That is, this method places a full distribution upon the test statistics (i.e., the second-stage mixture model) and allows for the posterior estimation of a probability of a true association conditional on the prior covariate structure. Because the hierarchical nature of the data—that is, SNPs within genes and genes within pathways—provides an opportunity to test from the "bottom up" in this analysis procedure, the method places more emphasis on tests of main effects or combinations of SNPs within a gene in comparison to SNP interactions across genes. By formalizing the joint distribution of all the test statistics, the prior beliefs in the relations between them, and the uncertainty of the model form, the parameter estimates and corresponding uncertainty intervals will better capture the dependency between terms. This, in turn, results in tests that more effectively reflect the evaluation of multiple factors. This is in contrast to more conventional approaches, such as the Bonferoni correction and controlling for false discovery rates, in which a uniform adjustment of the critical level is made across all p-values. By focusing on the posterior estimates for final inference, some of the multiple comparison pitfalls may be avoided. However, when relying on Bayes factors to gauge statistical significance, the influence of the prior structure is removed and the focus is solely on what the data tell us. Here, one must be careful when determining a cutoff level for declaring significance and should consider the number of comparisons made in deciding what is truly significant.

Summary

An overview has been presented of the analysis of numerous SNPs across multiple genes in a pathway focusing on the overall idea of incorporating prior knowledge via ontologies into a Bayesian hierarchical framework. The method presented is viewed as a unified approach by guiding statistical model selection with one's knowledge. In this framework, the method is based on the belief that polymorphisms, genes, and corresponding interactions vary in their biological plausibility and that by formally incorporating this differentiation into the statistical analysis, some of the difficulties in evaluating numerous factors may be lessened.

While there are many difficulties in pathway-based analyses, a pathway perspective has considerable promise. Many insights of relations and assumptions may be gained by properly representing one's knowledge of the underlying processes via ontologies and corresponding graphical representations. Furthermore, the formal incorporation of one's knowledge into the statistical framework can both guide the model search to more relevant models and allow interpretation of findings specifically in the context of one's knowledge base. Ultimately, confirmation of results by further studies is the key to valid conclusions in this area of research. However, this hierarchical model selection procedure with the incorporation of prior knowledge can help not only in identifying individual components but also in the characterization of the underlying complexity of a particular trait's variation.

Conclusions

1. The available knowledge of nicotine dependence arises largely from studies that model the independent association of candidate genes with outcome measures. Such studies often fail to reflect the complexity of interacting factors and discrete events that can influence smoking behavior and, therefore, may not provide a clear picture of biological mechanisms affecting nicotine dependence.
2. A promising approach to the study of nicotine dependence involves the use of prior biological knowledge about the relations between genotypic and phenotypic variables in a hierarchical modeling framework. This allows prior knowledge to aid in estimating specific genotypic effects and to guide a stochastic search over all possible statistical models.
3. The use of ontologies is a promising new direction for the elucidation of the genetic basis of nicotine dependence. An ontology is a construct or model that represents entities in both genotypic and phenotypic domains as well as their interrelations. The use of an ontology permits the modeling of hierarchical relationships by using directed acyclic graphs spanning genotypes and endophenotypes and phenotypes, while taking advantage of prior knowledge to quantify these relationships, making them amenable to computational analysis.
4. A study of nicotine metabolism that used data from the Northern California Twin Registry to examine the total clearance of nicotine and the *trans* 3′-hydroxycotinine to cotinine ratio, with the Nicotine Pharmacokinetics Ontology as a framework, showed a significant association between specific polymorphisms for *CYP2A6* and measured nicotine clearance levels as well as statistically significant results for single nucleotide polymorphism 4 within *UGT1A4*.
5. Hierarchical modeling combined with the use of an ontology defining relationships between constructs of interest represents a promising area for further research in studying a possible genetic basis for nicotine dependence as well as for understanding the interaction between genetics and social and environmental influences on tobacco use and dependence.

Appendix 12A. Estimation for the Hierarchical Model

A two-step estimation procedure is performed. First, for a given regression model, obtain the maximum likelihood estimates for $\hat{\beta}_v$ and $\hat{\text{var}}(\hat{\beta}_v)$ from a generalized linear model likelihood, $f(Y|X,\beta)$, and calculate the corresponding test statistic, t_v. Second, conditional on the set of test statistics, the contribution to the likelihood for each term in the second-stage model is the marginal distribution of t_v.

$$
\begin{aligned}
g\left(t_v \mid \underline{\mu},\underline{\pi},\sigma,\mathbf{Z}_\mu,\mathbf{Z}_\pi\right) &= \int a\left(t_v \mid \lambda_v\right) b\left(\lambda_v,\underline{\mu},\underline{\pi},\sigma,\mathbf{Z}_\mu,\mathbf{Z}_\pi\right) d\lambda_v \\
&= p_v \frac{a\left((t_v/\sqrt{1+\sigma^2})\Big/(\underline{\mu}'\mathbf{Z}_\mu\Big/\sqrt{1+\sigma^2})\right)}{\sqrt{1+\sigma^2}} + (1-p_v)a\left(t_v \mid 0\right)
\end{aligned}
$$

where $a()$ is the chi distribution given in equation (13) and $b()$ is the mixture distribution given in equation (14). The full log-likelihood for the second-stage model is then the marginal distribution summed over the entire set of test statistics,

$$
\sum_v \log\left(g\left(t_v \mid \underline{\mu},\underline{\pi},\sigma,\mathbf{Z}_\mu,\mathbf{Z}_\pi\right)\right)
$$

and maximized with respect to $\Theta = (\underline{\mu},\underline{\pi},\sigma)$.

Application of the Bayes formula results in expressions for the posterior of the probability of an association being true:

$$
\begin{aligned}
P_v &= \Pr(\lambda_v > 0 \mid t_v,\mathbf{Z}_\mu,\mathbf{Z}_\pi;\Theta) \\
&= \frac{1}{1+\dfrac{(1-p_v)}{p_v}\dfrac{a\left(t_v \mid 0\right)}{(1+\sigma^2)^{-1/2}a\left((t_v/\sqrt{1+\sigma^2})\Big/(\underline{\mu}'\mathbf{Z}_\mu\Big/\sqrt{1+\sigma^2})\right)}}
\end{aligned}
$$

and for the posterior magnitude of the association:

$$
\begin{aligned}
E_v &= E(\lambda_v \mid \lambda_v > 0, t_v, \mathbf{Z}_\mu;\Theta) \\
&= \frac{\sigma}{\sqrt{1+\sigma^2}} \frac{\sqrt{2/\pi}\exp\left(-\dfrac{(\sigma^2 t_v^2 + (\underline{\mu}'\mathbf{Z}_\mu)^2)}{2\sigma^2}\right) + \lambda_+\varphi(E_-)(2\Phi(\lambda_+)-1) + \lambda_-\varphi(E_+)(2\Phi(\lambda_-)-1)}{\varphi(E_+)+\varphi(E_-)}
\end{aligned}
$$

where

$$\lambda_{+} = (\underline{\mu}'\mathbf{Z}_{\mu} + \sigma^2 t_{\nu}) / \sigma\sqrt{1+\sigma^2}$$

$$\lambda_{-} = (\underline{\mu}'\mathbf{Z}_{\mu} - \sigma^2 t_{\nu}) / \sigma\sqrt{1+\sigma^2}$$

$$E_{+} = (t_{\nu} + \underline{\mu}'\mathbf{Z}_{\mu}) / \sqrt{1+\sigma^2}$$

$$E_{-} = (t_{\nu} - \underline{\mu}'\mathbf{Z}_{\mu}) / \sqrt{1+\sigma^2}$$

and Φ denotes the cumulative distribution function of a standard normal. Use a standard numerical maximization algorithm to maximize $\hat{\Theta}$. The estimated parameters $\hat{\Theta}$ are then substituted in the posterior expression to obtain $\hat{P}_{\nu}$ and $\hat{E}_{\nu}$.

Appendix 12B. Model Selection Algorithm

Assuming that the second-stage mixed model is independent of the regression model conditional on the test statistics, first define the posterior probability as

$$h\left(\underline{\gamma} \mid Y, \mathbf{X}\right) \propto \sum f\left(Y \mid \mathbf{X}, \underline{\beta}, \underline{\gamma}\right) \prod g\left(t_v \mid \mathbf{Z}_\mu, \mathbf{Z}_\pi; \Theta\right) q\left(\underline{\gamma}\right) d\gamma$$

where $f\left(Y \mid \mathbf{X}, \underline{\beta}, \underline{\gamma}\right)$ is the log-likelihood of the first-stage regression model. Because the model space is tremendous, one should not attempt to obtain a posterior estimation for the $\underline{\gamma}$s by integrating over all possible models. Instead, adopt an MCMC approach by using a Metropolis-Hastings algorithm.[113] Thus, during the iterations of the Markov chain, accept a new vector of $\underline{\gamma}$s, $\underline{\gamma}^*$ at iteration (i + 1) with probability

$$\alpha(\underline{\gamma}^i, \underline{\gamma}^*) = \min\left[1, \frac{h\left(\underline{\gamma}^* \mid Y, \mathbf{X}\right)}{h\left(\underline{\gamma}^i \mid Y, \mathbf{X}\right)} \frac{PD(\underline{\gamma}^i \mid \underline{\gamma}^*)}{PD(\underline{\gamma}^* \mid \underline{\gamma}^i)}\right]$$

Here, a proposal distribution *(PD)* is defined as a function of $\hat{P}_v$, the probability that a term is associated with the outcome. Specifically, *PD* is defined as

$$PD(\underline{\gamma}^* \mid \underline{\gamma}^i, \underline{\hat{P}}^i) = \prod_v \hat{P}_v^i \gamma_v^* + (1 - \hat{P}_v^i)(1 - \gamma_v^*)$$

where $\hat{P}_v^i = \Pr(\lambda_v > 0 \mid t_v, \gamma_v^i, \mathbf{Z}_\mu, \mathbf{Z}_\pi; \Theta)$. That is, the probability of a proposed vector of $\underline{\gamma}$s is dependent upon the probability that the terms are associated with the outcome given the vector of $\underline{\gamma}$s at iteration i.

The MCMC algorithm is

(continued on next page)

Initiate $\underline{\gamma}^0$ at $i = 0$

Repeat {

For iteration i

For all V,

Sample $\gamma_v^* \sim Bernoulli(\hat{P}_v^i)$

Obtain t_v^* from $f\left(Y \mid \mathbf{X}, \underline{t}^*, \underline{\gamma}^*\right)$ if $\gamma_v^* = 1$ and from $\chi_1(0)$ if $\gamma_v^* = 0$

Estimate $\hat{P}_v^*$ and $\hat{E}_v^*$ by maximizing $\sum_v g\left(t_v \mid \mathbf{Z}_\mu, \mathbf{Z}_\pi; \Theta\right)$

Calculate $PD(\underline{\gamma}^* \mid \underline{\gamma}^i, \underline{\hat{P}}^i)$ and $PD(\underline{\gamma}^i \mid \underline{\gamma}^*, \underline{\hat{P}}^*)$

Calculate $\alpha(\underline{\gamma}^t, \underline{\gamma}^*)$

Sample $u \sim U(0,1)$

If $u \leq \alpha(\underline{\gamma}^t, \underline{\gamma}^*)$

Then $\underline{\gamma}^{i+1} = \underline{\gamma}^*, \underline{\hat{P}}^{i+1} = \underline{\hat{P}}^*, \underline{\hat{E}}^{i+1} = \underline{\hat{E}}^*$

Else $\underline{\gamma}^{i+1} = \underline{\gamma}^i, \underline{\hat{P}}^{i+1} = \underline{\hat{P}}^i, \underline{\hat{E}}^{i+1} = \underline{\hat{E}}^i$

$i = i + 1$

}

References

1. Swan, G. E., N. L. Benowitz, P. Jacob 3rd, C. N. Lessov, R. F. Tyndale, K. Wilhelmsen, R. E. Krasnow, M. R. McElroy, S. E. Moore, and M. Wambach. 2004. Pharmacogenetics of nicotine metabolism in twins: Methods and procedures. *Twin Research* 7 (5): 435–48.
2. Swan, G. E., N. L. Benowitz, C. N. Lessov, P. Jacob 3rd, R. F. Tyndale, and K. Wilhelmsen. 2005. Nicotine metabolism: The impact of CYP2A6 on estimates of additive genetic influence. *Pharmacogenetics and Genomics* 15 (2): 115–25.
3. Dedobbeleer, N., F. Beland, A. P. Contandriopoulos, and M. Adrian. 2004. Gender and the social context of smoking behaviour. *Social Science and Medicine* 58 (1): 1–12.
4. Evans, R. I. 1976. Smoking in children: Developing a social psychological strategy of deterrence. *Preventive Medicine* 5 (1): 122–7.
5. Kandel, D. B., K. Yamaguchi, and K. Chen. 1992. Stages of progression in drug involvement from adolescence to adulthood: Further evidence for the gateway theory. *Journal of Studies on Alcohol* 53 (5): 447–57.
6. Shih, J. C., and R. F. Thompson. 1999. Monoamine oxidase in neuropsychiatry and behavior. *American Journal of Human Genetics* 65 (3): 593–98.
7. Brunner, H. G., M. Nelen, X. O. Breakefield, H. H. Ropers, and B. A. van Oost. 1993. Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science* 262 (5133): 578–80.
8. Cases, O., I. Seif, J. Grimsby, P. Gaspar, K. Chen, S. Pournin, U. Muller, et al. 1995. Aggressive behavior and altered amounts of brain serotonin and norepinephrine in mice lacking MAOA. *Science* 268 (5218): 1763–66.
9. Whitfield, J. B., D. Pang, K. K. Bucholz, P. A. Madden, A. C. Heath, D. J. Statham, and N. G. Martin. 2000. Monoamine oxidase: Associations with alcohol dependence, smoking and other measures of psychopathology. *Psychological Medicine* 30 (2): 443–54.
10. Brennan, P. 2002. Gene-environment interaction and aetiology of cancer: What does it mean and how can we measure it? *Carcinogenesis* 23 (3): 381–87.
11. Cordell, H. J. 2002. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11 (20): 2463–68.
12. Chatterjee, N., Z. Kalaylioglu, R. Moslehi, U. Peters, and S. Wacholder. 2006. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics* 79 (6): 1002–1016.
13. Kell, D. B. 2005. Metabolomics, machine learning and modelling: Towards an understanding of the language of cells. *Biochemical Society Transactions* 33 (Pt. 3): 520–24.
14. Thomas, C. E., and G. Ganji. 2006. Integration of genomic and metabonomic data in systems biology—are we 'there' yet? *Current Opinion in Drug Discovery & Development* 9 (1): 92–100.
15. Sellers, T. A., and J. R. Yates. 2003. Review of proteomics with applications to genetic epidemiology. *Genetic Epidemiology* 24 (2): 83–98.
16. Feng, Z., R. Prentice, and S. Srivastava. 2004. Research issues and strategies for genomic and proteomic biomarker discovery and validation: A statistical perspective. *Pharmacogenomics* 5 (6): 709–19.
17. Jones, P. A., and S. B. Baylin. 2002. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* 3 (6): 415–28.
18. Cusick, M. E., N. Klitgord, M. Vidal, and D. E. Hill. 2005. Interactome: Gateway into systems biology. *Human Molecular Genetics* 14 Spec No. 2: R171–R181.
19. Vidal, M. 2005. Interactome modeling. *FEBS Letters* 579 (8): 1834–38.
20. Wachi, S., K. Yoneda, and R. Wu. 2005. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21 (23): 4205–8.
21. Greenland, S. 1993. Methods for epidemiologic analyses of multiple exposures: A review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Statistics in Medicine* 12 (8): 717–36.
22. Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning.* New York: Springer.
23. Breiman, L. 2001. Random forests. *Machine Learning* 45 (1): 5–32.

24. Ruczinski, I., C. Kooperberg, and M. LeBlanc. 2003. Logic regression. *Journal of Computational and Graphical Statistics* 12 (3): 475–511.
25. Millstein, J., D. V. Conti, F. D. Gilliland, and W. J. Gauderman. 2006. A testing framework for identifying susceptibility genes in the presence of epistasis. *American Journal of Human Genetics* 78 (1): 15–27.
26. Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi, and I. Golani. 2001. Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research* 125 (1–2): 279–84.
27. Sabatti, C., S. Service, and N. Freimer. 2002. *False discovery rate in and correction for multiple comparisons in linkage disequilibrium genome screens,* Paper 2002010116. Los Angeles: Univ. of California Los Angeles. Department of Statistics. http://repositories.cdlib.org/uclastat/papers/2002010116.
28. Devlin, B., K. Roeder, and L. Wasserman. 2003. Analysis of multilocus models of association. *Genetic Epidemiology* 25 (1): 36–47.
29. Sabatti, C., S. Service, and N. Freimer. 2003. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* 164 (2): 829–33.
30. Whittemore, A. S. 2007. A Bayesian false discovery rate for multiple testing. *Journal of Applied Statistics* 34 (1): 1–9.
31. Thomas, D. C. 2005. The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiology, Biomarkers, & Prevention* 14 (3): 557–9.
32. Cortessis, V., and D. C. Thomas. 2003. Toxicokinetic genetics: An approach to gene-environment and gene-gene interactions in complex metabolic pathways. In *Mechanistic considerations in the molecular epidemiology of cancer,* ed. P. Bird, P. Boffetta, P. Buffler, and J. Rice, 127–50. Lyon, France: IARC Scientific Publications.
33. Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5 (2): 199–220.
34. McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models.* 2nd ed. Boca Raton, FL: CRC Press.
35. Schaid, D. J. 1996. General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology* 13 (5): 423–49.
36. Schaid, D. J. 2002. Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genetic Epidemiology* 23 (4): 426–43.
37. Zaykin, D. V., P. H. Westfall, S. S. Young, M. A. Karnoub, M. J. Wagner, and M. G. Ehm. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity* 53 (2): 79–91.
38. Stram, D. O., C. A. Haiman, J. N. Hirschhorn, D. Altshuler, L. N. Kolonel, B. E. Henderson, and M. C. Pike. 2003. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Human Heredity* 55 (1): 27–36.
39. Stram, D. O. 2005. Software for tag single nucleotide polymorphism selection. *Human Genomics* 2 (2): 144–51.
40. Cordell, H. J. 2006. Estimation and testing of genotype and haplotype effects in case-control studies: Comparison of weighted regression and multiple imputation procedures. *Genetic Epidemiology* 30 (3): 259–75.
41. Kraft, P., D. G. Cox, R. A. Paynter, D. Hunter, and I. De Vivo. 2005. Accounting for haplotype uncertainty in matched association studies: A comparison of simple and flexible techniques. *Genetic Epidemiology* 28 (3): 261–72.
42. Schaid, D. J. 2004. Evaluating associations of haplotypes with traits. *Genetic Epidemiology* 27 (4): 348–64.
43. Schaid, D. J. 2006. Power and sample size for testing associations of haplotypes with complex traits. *Annals of Human Genetics* 70 (Pt. 1): 116–30.
44. Conti, D. V., and W. J. Gauderman. 2004. SNPs, haplotypes, and model selection in a candidate gene region: The SIMPle analysis for multilocus data. *Genetic Epidemiology* 27 (4): 429–41.
45. Excoffier, L., and M. Slatkin. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* 12 (5): 921–7.
46. Robins, J. M., and S. Greenland. 1986. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology* 123 (3): 392–402.

47. Goodman, S. N. 1998. Multiple comparisons, explained. *American Journal of Epidemiology* 147 (9): 807–12.

48. Thompson, J. R. 1998. Invited commentary: Re: "Multiple comparisons and related issues in the interpretation of epidemiologic data." *American Journal of Epidemiology* 147 (9): 801–6.

49. Greenland, S. 2000. Principles of multilevel modelling. *International Journal of Epidemiology* 29 (1): 158–67.

50. Greenland, S. 2000. When should epidemiologic regressions use random coefficients? *Biometrics* 56 (3): 915–21.

51. Witte, J. S. 1997. Genetic analysis with hierarchical models. *Genetic Epidemiology* 14 (6): 1137–42.

52. Witte, J. S., and S. Greenland. 1996. Simulation study of hierarchical regression. *Statistics in Medicine* 15 (11): 1161–70.

53. Witte, J. S., S. Greenland, R. W. Haile, and C. L. Bird. 1994. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology* 5 (6): 612–21.

54. Wacholder, S., S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. Rothman. 2004. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute* 96 (6): 434–42.

55. Thomas, D. C., and D. G. Clayton. 2004. Betting odds and genetic associations. *Journal of the National Cancer Institute* 96 (6): 421–3.

56. Aragaki, C. C., S. Greenland, N. Probst-Hensch, and R. W. Haile. 1997. Hierarchical modeling of gene-environment interactions: Estimating NAT2 genotype-specific dietary effects on adenomatous polyps. *Cancer Epidemiology, Biomarkers & Prevention* 6 (5): 307–14.

57. Hung, R. J., P. Brennan, C. Malaveille, S. Porru, F. Donato, P. Boffetta, and J. S. Witte. 2004. Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer. *Cancer Epidemiology, Biomarkers & Prevention* 13 (6): 1013–21.

58. Conti, D. V., V. Cortessis, J. Molitor, and D. C. Thomas. 2003. Bayesian modeling of complex metabolic pathways. *Human Heredity* 56 (1-3): 83–93.

59. George, E. I., and R. E. McCulloch. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88 (423): 881–89.

60. George, E. I., and D. P. Foster. 2000. Calibration and empirical Bayes variable selection. *Biometrika* 87 (4): 731–47.

61. George, E. I., and R. E. McCulloch. 1995. Stochastic search variable selection. In *Markov chain Monte Carlo in practice,* ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, 203–14. London: Chapman and Hall.

62. Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. 2003. *WinBUGS user manual.* Version 1.4. Cambridge, UK: Univ. of Cambridge, Institute of Public Health. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf.

63. Chipman, H. 1996. Bayesian variable selection with related predictors. *Canadian Journal of Statistics* 24 (1): 17–36.

64. Chipman, H. A., E. I. George, and R. E. McCulloch. 2001. Managing multiple models. In *Artificial intelligence and statistics 2001,* ed. T. Jaakkola and T. Richardson. San Francisco: Morgan Kaufmann.

65. Lewinger, J. P., D. V. Conti, J. W. Baurley, T. J. Triche, and D. C. Thomas. 2007. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic Epidemiology* 31 (8): 871–82.

66. Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90 (430): 773–95.

67. Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294 (5547): 1719–23.

68. Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge, UK: Cambridge Univ. Press.

69. Fred Hutchinson Cancer Research Center. 2008. Sorting intolerant from tolerant. http://blocks.fhcrc.org/sift/SIFT.html (accessed December 19, 2008).

70. Harvard University. 2008. *PolyPhen:* Prediction of functional effect of human nsSNPs. http://genetics.bwh.harvard.edu/pph (accessed December 19, 2008).

71. SRI International. 2008. Evolutionary analysis of coding SNPS. http://www.pantherdb.org/tools/csnpScoreForm.jsp (accessed December 19, 2008).

72. Ng, P. C., and S. Henikoff. 2001. Predicting deleterious amino acid substitutions. *Genome Research* 11 (5): 863–74.

73. Sunyaev, S., V. Ramensky, I. Koch, W. Lathe 3rd, A. S. Kondrashov, and P. Bork. 2001. Prediction of deleterious human alleles. *Human Molecular Genetics* 10 (6): 591–7.

74. Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research* 13 (9): 2129–41.

75. Yue, P., E. Melamud, and J. Moult. 2006. SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics* 7:166.

76. Margulies, E. H., M. Blanchette, D. Haussler, and E. D. Green. 2003. Identification and characterization of multi-species conserved sequences. *Genome Research* 13 (12): 2507–18.

77. Research Collaboratory for Structural Bioinformatics. 2008. RCSB protein data bank. http://www.rcsb.org/pdb/explore.do?structureID=2FDU (accessed December 19, 2008).

78. Bader, G. D., M. P. Cary, and C. Sander. 2006. Pathguide: A pathway resource list. *Nucleic Acids Research* 34 (Database issue): D504–D506.

79. SRI International. 2008. HumanCyc database. http://biocyc.org/HUMAN/NEW-IMAGE?type=PATHWAY&object=PWY66-201 (accessed December 19, 2008).

80. SRI International. 2008. PANTHER classification system. http://www.pantherdb.org (accessed December 19, 2008).

81. Piper, M. E., D. E. McCarthy, and T. B. Baker. 2006. Assessing tobacco dependence: A guide to measure evaluation and selection. *Nicotine & Tobacco Research* 8 (3): 339–51.

82. Smith, B., W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. 2005. Relations in biomedical ontologies. *Genome Biology* 6 (5): R46.

83. Karp, P. D. 2001. Pathway databases: A case study in computational symbolic theories. *Science* 293 (5537): 2040–4.

84. Aranguren, M. E., S. Bechhofer, P. Lord, U. Sattler, and R. Stevens. 2007. Understanding and using the meaning of statements in a bio-ontology: Recasting the Gene Ontology in OWL. *BMC Bioinformatics* 8: 57.

85. Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25 (1): 25–29.

86. Diehl, A. D., J. A. Lee, R. H. Scheuermann, and J. A. Blake. 2007. Ontology development for biological systems: Immunology. *Bioinformatics* 23 (7): 913–15.

87. Gkoutos, G. V., E. C. Green, A. M. Mallon, J. M. Hancock, and D. Davidson. 2004. Building mouse phenotype ontologies. *Pacific Symposium on Biocomputing:* 178–89.

88. Yu, A. C. 2006. Methods in biomedical ontology. *Journal Biomedical Informatics* 39 (3): 252–66.

89. Noy, N. F., and D. L. McGuinness. 2001. *Ontology development 101: A guide to creating your first ontology.* Technical Report KSL-01-05. Palo Alto, CA: Stanford Univ., Stanford Knowledge Systems Laboratory. http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf.

90. Gruninger, M., and M. S. Fox. 1995. Workshop on basic ontological issues in knowledge sharing: Methodology for the design and evaluation of ontologies. In *1995 International Joint Conference on Artificial Intelligence,* ed. D. Skuce, 6.1–6.10. San Francisco: Morgan Kaufman.

91. OBO Foundry. 2008. The open biomedical ontologies. http://www.obofoundry.org (accessed December 19, 2008).

92. Distelhorst, G., V. Srivastava, C. Rosse, and J. F. Brinkley. 2003. A prototype natural language interface to a large complex knowledge base, the Foundational Model of Anatomy. *AMIA Annual Symposium Proceedings:* 200–204.

93. *Nucleic Acids Research.* 2006. The Gene Ontology (GO) project in 2006. *Nucleic Acids Research* 34 (Database issue): D322–6.

94. Kushida, T., T. Takagi, and K. I. Fukuda. 2006. Event ontology: A pathway-centric ontology for biological processes. *Pacific Symposium on Biocomputing:* 152–63.

95. Eilbeck, K., S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. 2005. The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* 6 (5): R44.

96. W3C. 2008. OWL web ontology language. http://www.w3.org/TR/owl-features (accessed December 19, 2008).
97. Berkeley Bioinformatics and Ontologies Project. 2008. The OBO ontology editor. http://oboedit.org (accessed December 19, 2008).
98. Noy, N. F., M. Crubezy, R. W. Fergerson, H. Knublauch, S. W. Tu, J. Vendetti, and M. A. Musen. 2003. Protege-2000: An open-source ontology-development and knowledge-acquisition environment. *AMIA Annual Symposium Proceedings:* 953.
99. Luciano, J. S. 2005. PAX of mind for pathway researchers. *Drug Discovery Today* 10 (13): 937–42.
100. Karp, P. D., C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas. 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research* 33 (19): 6083–9.
101. Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. 2006. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Research* 34 (Database issue): D354–7.
102. Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, et al. 2005. Reactome: A knowledge base of biological pathways. *Nucleic Acids Research* 33 (Database issue): D428–32.
103. Mi, H., N. Guo, A. Kejariwal, and P. D. Thomas. 2007. PANTHER version 6: Protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research* 35 (Database issue): D247–52.
104. Hucka, M., A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, et al. 2003. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* 19 (4): 524–31.
105. Kitano, H., A. Funahashi, Y. Matsuoka, and K. Oda. 2005. Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology* 23 (8): 961–6.
106. Rubin, D. L., S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, et al. 2006. National Center for Biomedical Ontology: Advancing biomedicine through structured organization of scientific knowledge. *OMICS* 10 (2): 185–98.
107. SourceForge. 2008. Open source software. http://sourceforge.net (accessed December 19, 2008).
108. Benowitz, N. L., G. E. Swan, P. Jacob 3rd, C. N. Lessov-Schlaggar, and R. F. Tyndale. 2006. CYP2A6 genotype and the metabolism and disposition kinetics of nicotine. *Clinical Pharmacology and Therapeutics* 80 (5): 457–67.
109. Dempsey, D., P. Tutka, P. Jacob 3rd, F. Allen, K. Schoedel, R. F. Tyndale, and N. L. Benowitz. 2004. Nicotine metabolite ratio as an index of cytochrome P450 2A6 metabolic activity. *Clinical Pharmacology and Therapeutics* 76 (1): 64–72.
110. Berkeley Drosophila Genome Center. 2008. Phenotype syntax. http://www.fruitfly.org/~cjm/obd/pheno-syntax.html (accessed December 19, 2008).
111. Hukkanen, J., P. Jacob 3rd, and N. L. Benowitz. 2005. Metabolism and disposition kinetics of nicotine. *Pharmacological Reviews* 57 (1): 79–115.
112. R Development Core Team. 2003. *The R project for statistical computing.* Vienna, AU: R Foundation for Statistical Computing. http://www.r-project.org.
113. Gilks, W. R., S. Richardson, and D. Spiegelhalter, ed. 1996. *Markov chain Monte Carlo in practice.* London: Chapman and Hall.
114. Srivastava, V. K., and D. C. Hill. 1975. Thiocyanate ion formation in rapeseed meals. *Canadian Journal of Biochemistry* 53 (5): 630–33.
115. Oscarson, M., R. A. McLellan, V. Asp, M. Ledesma, M. L. Bernal Ruiz, B. Sinues, A. Rautio, and M. Ingelman-Sundberg. 2002. Characterization of a novel CYP2A7/CYP2A6 hybrid allele (CYP2A6*12) that causes reduced CYP2A6 activity. *Human Mutation* 20 (4): 275–83.
116. Yamano, S., J. Tatsuno, and F. J. Gonzalez. 1990. The CYP2A3 gene product catalyzes coumarin 7-hydroxylation in human liver microsomes. *Biochemistry* 29 (5): 1322–9.
117. Nunoya, K., T. Yokoi, K. Kimura, K. Inoue, T. Kodama, M. Funayama, K. Nagashima, et al. 1998. A new deleted allele in the human cytochrome P450 2A6 (CYP2A6) gene found in individuals showing poor

metabolic capacity to coumarin and (+)-cis-3,5-dimethyl-2-(3-pyridyl)thiazolidin-4-one hydrochloride (SM-12502). *Pharmacogenetics* 8 (3): 239–49.

118. Wakefield, J. 2007. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics* 81 (2): 208–27.

Part 6

Future Directions

This final part explores future directions for genetic studies of nicotine dependence, taking into account social and environmental influences as well as complex G×E interactions. It presents chapter-specific and cross-cutting recommendations for future research in tobacco genetics.