Published in final edited form as:

Int J Data Min Bioinform. 2015 ; 12(2): 129–143.

Discovery of Phenotypic Networks from Genotypic Association Studies with Application to Obesity

Christine W. Duarte¹, Yann C. Klimentidis¹, Jacqueline J. Harris¹, Michelle Cardel², and José R. Fernández²

¹ Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, 35294

² Department of Nutrition Sciences, University of Alabama at Birmingham, AL, 35294

Abstract

Genome-wide Association Studies (GWAS) have resulted in many discovered risk variants for several obesity-related traits. However, before clinical relevance of these discoveries can be achieved, molecular or physiological mechanisms of these risk variants needs to be discovered. One strategy is to perform data mining of phenotypically-rich data sources such as those present in dbGAP (database of Genotypes and Phenotypes) for hypothesis generation. Here we propose a technique that combines the power of existing Bayesian Network (BN) learning algorithms with the statistical rigor of Structural Equation Modeling (SEM) to produce an overall phenotypic network discovery system with optimal properties. We illustrate our method using the analysis of a candidate SNP data set from the AMERICO sample, a multi-ethnic cross-sectional cohort of roughly three hundred children with detailed obesity-related phenotypes. We demonstrate our approach by showing genetic mechanisms for three obesity-related SNPs.

Keywords

genetic networks; GWAS; Bayesian Networks; Structural Equation Modeling; obesity

INTRODUCTION

There is a growing epidemic of obesity in the United States that is a major concern for public health. Obesity is a risk factor for cardiovascular disease, diabetes, cancer, and many other diseases that are leading causes of death, disability, and health care expenditure in the United States. Tremendous public investment has been made into finding genetic determinants of obesity, most recently in the form of GWAS (genome-wide association studies). GWAS studies have resulted in the discovery of many associated genetic loci; for instance, the GIANT consortium has discovered or validated a total of 32 SNPs associated with body mass index (BMI) (Hebebrand, Volckmar et al. 2010; Speliotes, Willer et al. 2010).

While many successes in the genetics of obesity have occurred, the overall assessment of the contribution of GWAS remains controversial, largely due to the lack of translation of basic scientific discoveries into improved health outcomes for patients. There are several possible

reasons for this failure. One is that the overall explained variance of the discovered loci is not high. For instance, only 1-2% of the variance of BMI is explained by the 32 loci found in the GIANT study (Hebebrand, Volckmar et al. 2010), and thus most of the genetic variation in obesity remains undiscovered. However, there are examples in many GWAS studies in which discovered genetic variants, even of small effect, yield new insight into the pathophysiology of disease. For instance, in Type II diabetes (T2D), while only 10% of the disease heritability is accounted for by the set of discovered genetic variants, novel pathways have been discovered with previously-unknown roles in T2D, and the role of β -cell dysfunction in T2D has been highlighted as a result of these studies (Billings and Florez 2010).

For GWAS of BMI, some of the newly-discovered variants are in genes expressed in the central nervous system, and key hypothalamic pathways of energy balance are potentially implicated (Hebebrand, Volckmar et al. 2010). However, many of the discovered variants are in or near genes that are poorly-understood with unknown genetic mechanisms. In order to translate discovered genetic variants into new therapies or prevention strategies, first molecular and/or physiological mechanisms of each variant needs to be discovered. However, it is difficult to know how to explore potential genetic risk mechanisms, especially when little is known about a gene. One potential strategy is to use data mining of existing data sets in which detailed phenotypic data is available, and use machine learning and/or statistical modeling of this data for hypothesis generation. This strategy holds much promise because of the many large, publicly-available data sets available in dbGAP.

However, the set of statistical and computational methods for data analysis is limited. While many powerful and robust methods for GWAS analysis have been developed, the vast majority of methods are unigenic and univariate, as required for the high dimensionality of GWAS data. Notable exceptions include methods for multi-trait analysis (Ferreira and Purcell 2009; Yang, Wu et al. 2010), epistatic methods (Wan, Yang et al. 2010), and pathway-based methods for collapsing association signals for genes or pathways for improved power (Wang, Li et al. 2010). However, none of these methods are optimized for finding detailed mechanisms of action for genetic variants that likely involve the combined influence of multiple, interacting phenotypes.

Bayesian Networks (BN) is a machine learning approach for discovering relationships among a set of variables from an observational data set. The goal of BN learning algorithms is to find a Directed Acyclic Graph (DAG) which represents the true probability distribution for a set of random variables. In this graph the nodes are the random variables, $V_1, V_2, \dots, V_n$, and the edges signify probabilistic relationships. Bayesian Networks satisfy the Markov condition which states that each node is independent of its non-descendants given its parents,

$$P(V_1, V_2, \dots, V_n) = \prod_{i=1}^n P(V_i | \text{Parents}(V_i))$$

(Peter Spirtes 2000), where the parents of a node are defined as those nodes with edges directed into the node, and the non-descendants are defined as nodes for which there is not a directed path from the original node. See Figure 1 for an illustration of two DAGs and the resulting probability distribution associated with each graph. Networks can be learned for a data set using two different approaches. One is the score-based approach in which a quantity such as the likelihood or the posterior probability of a model is maximized after searching over a set of potential networks. A second approach is the conditional independence approach in which the conditional independence relations found in the data are directly used to infer network structure (for instance in the PC algorithm, (Peter Spirtes 2000)). The advantages of BN learning algorithms are that they are powerful, they produce graphical models that are intuitive and interpretable, and several optimized algorithms and software programs are currently available. The disadvantages are that they are computationally intensive (NP-Hard), and existing algorithms are not suitable for large data sets without drastic modifications. An additional disadvantage is that the statistical interpretation and level of confidence in whole networks or network features can be difficult to obtain from existing algorithms.

Structural Equation Modeling (SEM), and path analysis in particular, has the same overall goal as BN, which is to learn the causal relationships among a set of random variables. However, SEM methods are more focused on assessment of the fit of proposed models rather than searching for or proposing models for a given data set. SEM has several advantages including a sound theoretical basis, several well-developed methods and software platforms (such as MPLUS, LISREL, AMOS, and others), accurate model evaluation statistics such as the chi-squared goodness of fit statistic, the root mean squared error of approximation (RMSEA), and many others, and parameter estimates and standard errors for all model parameters that have appropriate Type I errors under the specified model assumptions. SEM also has disadvantages that include a lack of flexibility with regard to variables and models that can be fit, strong distributional assumptions, and a limited number of variables to model before convergence can fail. Therefore, SEM is not appropriate (by design) for data mining.

To overcome the limitations of BN and SEM when used by themselves, namely, the lack of statistical inference properties for BN, and the inappropriateness of SEM for data mining, we thus propose a hybrid approach that overcomes these limitations to create an overall method that is optimal for our purposes. The BN aspects that will be emphasized are the existence of powerful and efficient algorithms to explore model space. The SEM aspects that will be emphasized are the ability to evaluate model uncertainty, provide confidence intervals for model parameters, and in general perform statistical inference of BN-proposed models. We thus propose a two-part strategy in which BN algorithms are used to propose a set of models to test, and SEM modeling is used to evaluate and select/prioritize models for further study.

MATERIALS AND METHODS

Description of Data Set for Sample Application (the Americo Study)

A total of 294 children, aged 7 to 12 years (53% male), were recruited as part of a cross-sectional cohort study examining population differences in metabolic phenotypes among healthy children (no major illnesses or medical diagnoses). Race/ethnicity was determined by the parents of the subjects who could classify their children into either of these categories: African American (AA; n=96), Hispanic American (HA; n=78), European American (EA; n=114), or Bi-racial (n=6). All children were pubertal stage 3 as assessed by a pediatrician according to the criteria of Marshall and Tanner (Marshall and Tanner 1968). Informed assent and consent were obtained from children and parents respectively, as approved and regulated by the University of Alabama at Birmingham Institutional Review Board. All measurements were taken between 2004 and 2008 at the University of Alabama at Birmingham General Clinical Research Center (GCRC) and the Department of Nutrition Sciences. Phenotypes consisted of anthropometric measurements, insulin, glucose, and lipid measurements, blood pressure, body composition assessed by dual-energy x-ray absorptiometry (DXA), and physical activity assessed by accelerometer. Details of the study protocol and phenotype collection can be found elsewhere (Casazza, Dulin-Keita et al. 2009; Casazza, Gower et al. 2009).

Genotyping and Univariate Association Analysis

96 SNPs with validated associations from prior GWAS studies for BMI, hypertension, Type II Diabetes, and other obesity-related traits were selected for genotyping in the Americo sample (Meyre, Delplanque et al. 2009; Rivadeneira, Styrkarsdottir et al. 2009; Thorleifsson, Walters et al. 2009; Willer, Speliotes et al. 2009; Speliotes, Willer et al. 2010; Teslovich, Musunuru et al. 2010; Voight, Scott et al. 2010; Franceschini, Reiner et al. 2011). We present the association results and discovered networks for three representative SNPs in this study (see Tables 1 and 2 and Figure 2). DNA was obtained from all study participants and genotyped using Illumina Golden Gate at the UAB Heflin Genotyping Core. Each SNP was individually tested using an additive model for association with each trait, adjusting for age, Tanner stage, sex, racial/ethnic group, and height² (for total body fat and total lean mass). The overall data set consists of a set of obesity-related phenotypes (see Table S4), and 96 SNPs that are validated genetic variants associated with obesity-related traits in prior GWAS studies (see Table S5).

Specific Algorithm

The overall goal of the method is to take each SNP and set of associated phenotypes, and to propose and evaluate a network model for each SNP. The Bayesian Network algorithm that we have chosen for model proposal is called DEAL and is implemented in the R programming language (<http://cran.r-project.org/web/packages/deal/index.html>). DEAL is appropriate for analyzing joint distributions of continuous and categorical variables. It implements a Bayesian method with conjugate updating of network parameters for conditionally Gaussian networks (Bøttcher and Dethlefsen 2003). The SEM modeling software chosen for model assessment is MPLUS (www.statmodel.com). We use MLM estimation (i.e. the Satorra-Bentler chi-squared test (Chou, Bentler et al. 1991)) which is

designed to be robust in the presence of normality deviations in continuous traits. The requirements for an acceptable model fit in MPLUS included a non-significant chi-squared goodness of fit statistic, an RMSEA < 0.05 (root mean squared error of approximation), an SRMR < 0.05 (standardized root mean residual), an CFI > 0.95 (comparative fit index), and the majority of edges with significant coefficients ($p < 0.05$ using a Wald test) for most or all of the edges. To summarize, the hybrid Bayesian Network Structural Equation Modeling (BN/SEM) approach that we have implemented consists of the following steps:

1. For each SNP, find the set of associated traits at a pre-determined p -value threshold (after correcting for covariates).
2. For the SNP, covariates, and associated traits from the previous step, discover the highest-scoring network using DEAL (see details below).
3. Export DEAL network to MPLUS and fit network and obtain model diagnostics.
4. If model fit is appropriate, stop. Otherwise, make modifications to network until model fit is adequate (see details for acceptable model fit, previous paragraph).

The entire procedure including data import, univariate association analysis, Bayesian Network discovery using DEAL, export of discovered network to MPLUS, and MPLUS model evaluation, has been implemented in a custom R program which is available upon request. Final models from MPLUS were visualized using Cytoscape (<http://www.cytoscape.org>).

Bayesian Network Algorithm Details

DEAL is a package for learning Bayesian Networks in the R programming language. DEAL uses a Bayesian framework for learning networks of discrete and continuous variables using a conditional Gaussian likelihood with conjugate updating implemented used a heuristic search strategy. Banlists can be used to disallow edges directed into exogenous variables (such as the SNP and the covariates). The “learn” and “autosearch” functions are used for initializing and searching for the best network, respectively, and default prior distributions are used for all variables.

SEM Modeling Details

Structural models were fit in MPLUS (www.statmodel.com) using MLM estimation (i.e. the Satorra-Bentler chi-squared test) which is designed to be robust in the presence of normality deviations in the modeling of continuous traits. All of the endogenous variables were continuous. Covariates that were categorical were coded as dummy variables. Requirements for an acceptable model fit included: non-significant chi-squared goodness of fit statistic at a level of 0.05, RMSEA (root mean square error of approximation) < 0.06 , CFI (comparative fit index) > 0.95 , SRMR (standardized root mean residual) < 0.08 , and model parameters significant at a $p < 0.1$ for most or all of the edges.

If the DEAL-proposed network did not achieve an adequate fit according to the previously-stated criteria, then a series of steps would be taken to modify the model which consisted of 1) adding edges from each covariate to each endogenous variable and then removing covariate edges which are not significant at $p < 0.1$; 2) iteratively adding, removing, or

changing the direction of edges between endogenous variables, and checking for an improvement in model fit. Usually step 1) was sufficient to achieve an adequate model fit (adding covariates to the model), but occasionally step 2) was required to achieve an adequate model fit. Typically, the structure proposed by DEAL was for the most part preserved in the final model.

RESULTS

Here we present the network results for three representative genetic variants: one associated variant for each of three different GWAS traits: Type II Diabetes (T2D), body mass index (BMI), and blood pressure/hypertension (BP/HT). Table 1 shows the univariate association results for all traits with $p < 0.05$ for the three specific SNPs (rs4402960 in the gene IGF2BP2, rs2681492 in the gene ATP2B1, and rs7561317 in the gene TMEM18). DEAL-proposed networks for each of the SNPs were evaluated in MPLUS, and minor modifications were made to achieve acceptable fits. Figure 2 shows the resulting final networks for all three SNPs.

The biological interpretation of the first network for the SNP rs4402960 in the gene IGF2BP2 for Type II Diabetes (T2D) is that independent mechanisms of genetic risk affect the two phenotypes, insulin sensitivity (si) and total lean mass (dtotlean). The interpretation of the second network for the SNP rs2681492 in the gene ATP2B1 for blood pressure/hypertension (BP/HT) is that the genetic effect on the target trait, blood pressure, is potentially mediated by activity level. Finally, the interpretation for the third network for the SNP rs7561317 in TMEM18 for BMI (body mass index) is that a mediation model involving abdominal adipose tissue as well as total lean mass can explain the genetic effect of the SNP on BMI. Overall we show that we can obtain biologically-plausible physiological networks for a representative set of genetic variants and obesity-related traits.

DISCUSSION

We demonstrate a new computational system that proposes physiological mechanisms of genetic risk for associated variants discovered in GWAS studies using the analysis of phenotypically-rich data sets. Our system combines the power of Bayesian Network discovery algorithms for searching over the large set of potential networks with the statistical rigor of Structural Equation Modeling for assessing adequate model fit. We demonstrate our approach using preliminary results from a candidate SNP study in the Americo data set, a multi-ethnic cross-sectional study of children with detailed obesity-related phenotypes, giving three discovered genetic mechanisms as an illustration of our approach. Our future work on this computational system will include a detailed simulation study to assess the power and sample size requirements for networks of varying complexity, as well as analysis using larger data sets such as MESA, the Multi-Ethnic Study of Atherosclerosis (MESA) SHARe, which is publicly-available through dbGAP (<http://www.ncbi.nlm.nih.gov/gap>).

It is stressed here that the intention of this method is not for the initial analysis of genome-wide association studies for discovery of new associations, but rather for the detailed

characterization and modeling of existing associations using a phenomic data set. Thus the computational system is designed to handle hundreds rather than thousands or millions of genetic variants, since a precondition for inclusion of the genetic variant is prior association with a trait of interest. The method then builds a multivariate model of from tens to hundreds of phenotypes for each genetic variant, to probe for specific physiological mechanisms of genetic risk. Extension of the method to multigenic models is a future goal.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors would like to thank the Americo study participants and families, and acknowledge the following NIH grants: R01-DK067426, NIH T32-HI007457, 5P30DK056336-09, CA47888, and P60-DK079626. The opinions expressed are those of the authors and not necessarily those of the NIH or any other organization with which the authors are affiliated.

REFERENCES

- Billings LK, Florez JC. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci.* 2010; 1212:59–77. [PubMed: 21091714]
- Böttcher, SG.; Dethlefsen, C. Learning Bayesian Networks with R.. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria. 2003.
- Casazza K, Dulin-Keita A, et al. Differential influence of diet and physical activity on components of metabolic syndrome in a multiethnic sample of children. *J Am Diet Assoc.* 2009; 109(2):236–244. [PubMed: 19167950]
- Casazza K, Gower BA, et al. Physical fitness, activity, and insulin dynamics in early pubertal children. *Pediatr Exerc Sci.* 2009; 21(1):63–76. [PubMed: 19411712]
- Chou CP, Bentler PM, et al. Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *Br J Math Stat Psychol* 44. 1991:347–357.
- Ferreira MAR, Purcell SM. A multivariate test of association. *Bioinformatics.* 2009; 25(1):132–133. [PubMed: 19019849]
- Franceschini N, Reiner AP, et al. Recent findings in the genetics of blood pressure and hypertension traits. *Am J Hypertens.* 2011; 24(4):392–400. [PubMed: 20948529]
- Hebebrand J, Volckmar AL, et al. Chipping away the 'missing heritability': GIANT steps forward in the molecular elucidation of obesity - but still lots to go. *Obes Facts.* 2010; 3(5):294–303. [PubMed: 20975295]
- Marshall WA, Tanner JM. Growth and physiological development during adolescence. *Annu Rev Med.* 1968; 19:283–300. [PubMed: 4297619]
- Meyre D, Delplanque J, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet.* 2009; 41(2):157–159. [PubMed: 19151714]
- Peter Spirtes, CG.; Scheines, Richard. Causation, Prediction, and Search. The MIT Press; 2000.
- Rivadeneira F, Styrkarsdottir U, et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet.* 2009; 41(11):1199–1206. [PubMed: 19801982]
- Spiotes EK, Willer CJ, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet.* 2010; 42(11):937–948. [PubMed: 20935630]
- Teslovich TM, Musunuru K, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature.* 2010; 466(7307):707–713. [PubMed: 20686565]

- Thorleifsson G, Walters GB, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet.* 2009; 41(1):18–24. [PubMed: 19079260]
- Voight BF, Scott LJ, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet.* 2010; 42(7):579–589. [PubMed: 20581827]
- Wan X, Yang C, et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet.* 2010; 87(3):325–340. [PubMed: 20817139]
- Wang K, Li M, et al. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet.* 2010; 11(12):843–854. [PubMed: 21085203]
- Willer CJ, Speliotes EK, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet.* 2009; 41(1):25–34. [PubMed: 19079261]
- Yang QO, Wu HS, et al. Analyze Multivariate Phenotypes in Genetic Association Studies by Combining Univariate Association Tests. *Genetic Epidemiology.* 2010; 34(5):444–454. [PubMed: 20583287]

Figure 1.

Example of the Markov Condition for 2 Bayesian Networks, represented by Model 1 and Model 2, each containing nodes A, B, and C. The joint probability distribution of these variables can be decomposed into the conditional distribution of each variable given its parents. The two graphs shown result in two statistically-distinguishable probability distributions, each shown below each graph.

Figure 2.

Discovered networks for three genetic variants associated with obesity-related traits in the Americo study. *A* is the network for the SNP rs4402960 in the gene IGF2BP2, a GWAS discovery for Type II Diabetes (T2D). dtotlean is total lean mass normalized by height squared, and si is insulin sensitivity. *B* is the network for the SNP for rs2681492 in the gene ATP2B1, a GWAS discovery for blood pressure and hypertension (BP/HT). acwdmod is a measure of physical activity, and AVE_DIA is average diastolic blood pressure. *C* is the network for the SNP rs7561317 in TMEM18, a GWAS discovery for BMI (body mass

index). iaat and saat are intra-abdominal and subcutaneous abdominal adipose tissue, respectively. The covariates are EA and AA (for European American and African American ethnicity indicator variables), sex, age, and Tanner stage.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Univariate association results for each phenotype studied with three SNPs associated with obesity-related traits that showed nominal significance ($p < 0.05$). Discovered networks are shown in Figure 2.

SNP	Gene	Trait	Estimate	p-value
rs4402960	IGF2BP2	dtotlean (total lean mass)	−0.03	0.0032
rs4402960	IGF2BP2	si (insulin sensitivity)	0.96	0.0033
rs2681492	ATP2B1	AVE_DIA (avg. diastolic blood pressure)	1.68	0.0442
rs2681492	ATP2B1	acwdmod (moderate activity min./day)	8.70	0.0472
rs7561317	TMEM18	bmi (body mass index)	−0.73	0.0183
rs7561317	TMEM18	dtotlean (total lean mass)	−0.03	0.0173
rs7561317	TMEM18	iaat (intra-abdominal adipose tissue)	−6.46	0.0265
rs7561317	TMEM18	saat (subcutaneous adipose tissue)	−19.27	0.0427

Table 2

Model fit statistics (from MPLUS) for the networks shown in Figure 2. Network A is for the SNP rs4402960 in the gene IGF2BP2 associated with Type II Diabetes (T2D); network B is for the SNP rs2681492 in the gene ATP2B1 associated with blood pressure and hypertension (BP/HT); and network C is for the SNP rs7561317 in the gene TMEM18 associated with body mass index (BMI).

Network	Statistic	Value	df	p-value
A	Chi-Sq	0.694	2	0.7066
	RMSEA	0.000		
	CFI	1.000		
	SRMR	0.006		
B	Chi-Sq	1.092	3	0.7776
	RMSEA	0.000		
	CFI	1.000		
	SRMR	0.014		
C	Chi-Sq	12.10	15	0.6711
	RMSEA	0.000		
	CFI	1.000		
	SRMR	0.021		