Published in final edited form as:

Proc Ohio Collab Conf Bioinform. 2009 June 17; : 3–8. doi:10.1109/OCCBIO.2009.18.

Modeling Genetic and Environmental Factors in Biological Systems Using Structural Equation Modeling: An Application to Energy Balance

Nora L. Nock*, Li Li, and Robert C. Elston

Department of Epidemiology and Biostatistics Center for Transdisciplinary Research on Energetics and Cancer Case Western Reserve University Cleveland, OH 44106 U.S.A.

Abstract

To improve our understanding of the role(s) that genes and environmental factors play in a complex disease, we need statistical approaches that model multiple factors simultaneously in a hierarchical manner that aims to reflect the underlying biological system(s). We present an approach that models genes as latent constructs, defined by multiple variants (single nucleotide polymorphisms, SNPs) within each gene, using the multivariate statistical framework of structural equation modeling (SEM) to model multiple, putative genetic and environmental factors involved in energy imbalance ('obesity') using subjects from a colon polyp case-control study. We found that modeling constructs for the leptin receptor (LEPR) gene (defined by SNPs rs1137100, rs1137101, rs1805096, rs6588147) and the fat mass-and-obesity-associated (FTO) gene (defined by SNPs rs9939609, rs1421085, rs8044769) together with demographic (age, race, gender), physical activity, diet and sleep variables increased the strength of the association ($\beta_{\text{std}} = -0.13 \pm 0.06$; $p=0.03$) between the FTO and obesity constructs compared to that observed in a reduced model with only the FTO and LEPR constructs and demographic variables ($\beta_{\text{std}} = -0.05 \pm 0.03$; $p=0.08$). Several indirect paths, including an association between the LEPR and physical activity constructs ($\beta_{\text{std}} = -0.15 \pm 0.04$; $p=0.01$), were found. Interestingly, removing FTO revealed a marginal association between the LEPR and obesity constructs ($\beta_{\text{std}} = 0.24 \pm 0.14$; $p=0.09$), which was not present when FTO was in the model. These results illustrate the importance of modeling multiple relevant genes and other factors in the same model, which is a major strength of this approach. Moreover, our latent gene construct approach exploits the correlation structure between SNPs while capturing overall effects of variation in that gene, which will enable better utilization of candidate gene and genome-wide SNP array data.

Keywords

biological systems; structural equation modeling; FTO; LEPR; obesity

I. Introduction

We purport that the statistical framework of structural equation modeling (SEM) is valuable for mathematically describing the hierarchical relationships between genetic and environmental factors involved in the underlying biological systems driving complex diseases. The origins of SEM come from path analysis, which was initially described by Sewall Wright in 1921 [1]. However, SEM became more widely used, particularly in the psychometric and econometric fields, in the late 1970s to early 1980s when a program called

*Corresponding Author: nln@case.edu.

LISREL was developed which provided an efficient way of solving systems of linear equations [2,3].

In particular, we propose to use a latent gene construct SEM approach whereby we let a latent (not directly observable) variable represent the overall variation in a gene, which we formally describe by multiple SNPs in that gene [4]. Here, we describe the latent gene construct approach in more methodological detail and illustrate this approach by simultaneously modeling multiple genes together with physical activity, dietary, sleep and demographic variables putatively involved in the development of energy imbalance ('obesity') in a colon polyp case-control study data set.

Specifically, we examine potential effects of two key genes, the leptin receptor (LEPR) gene and the fat mass-and-obesity-associated (FTO) gene, since both have SNPs that were previously shown to be associated with waist circumference and/or body mass index (BMI) [5,6,7], which are surrogate measures of energy balance; however, associations between LEPR polymorphisms and obesity appear less inconsistent across studies [8]. Although it is well established that leptin acts on its receptors, which are located in the hypothalamus, to regulate energy intake and energy expenditure, the exact function of the FTO gene is unknown. Interestingly, the FTO (or 'FATSO') gene was first cloned through the fused toes mouse mutation [9], but recent studies indicate that FTO may alter satiety mechanisms [10,11]. Furthermore, although the relationship between obesity ($\text{BMI} \geq 30 \text{ kg/m}^2$) and colon cancer has been well established [12], the potential association between the FTO and LEPR genes on components of energy balance, and between energy imbalance and colon polyps directly, has not been well studied. Therefore, in our model, we evaluate direct relationships between these genes and measures of physical activity, dietary fat and sleep, which may play a key role in energy imbalance ('obesity').

Through this application, we aim to show that our latent gene construct SEM approach is advantageous in that it capitalizes on the substantial level of correlation between SNPs in the same gene, has the ability to estimate overall gene effects, improves mathematical representation of the underlying biological system(s) and provides improved control of confounding compared to standard regression methods.

II. Methods

A. Study Population

Patients coming to the University Hospitals Health System (UHHS) for routine colonoscopy screening were recruited prospectively from January, 2006 to August, 2008. Potential subjects were sent a letter introducing the study and were contacted by study personnel approximately one week later to complete a preliminary screening questionnaire. Patients with a known personal history of cancer or colorectal adenomatous polyps, or a known family history of hereditary non-polyposis colon cancer (HNPCC) or familial adenomatous polyps (FAP), were not eligible to participate. For patients who consented to be in the study, medical records and pathology reports from the colonoscopy were retrieved. Patients with histologically confirmed rectal polyps or colorectal cancer were excluded from this study. Patients with histologically confirmed colon adenomas were classified as cases. Patients who did not have polyps were classified as controls. The study was approved by the UHHS Institutional Review Board.

B. Body Composition, Diet, Physical Activity and Sleep

Eligible subjects who consented to participate in the study had their weight, height, and waist circumference measured (without shoes and in light clothes) by a research nurse at the

time of colonoscopy. Waist circumference was recorded in centimeters (cm) and body mass index (BMI) was calculated from the subject's weight (kg) divided by height squared (m²).

Subjects were required to complete a computer-aided personal interview on lifestyle factors at the time of colonoscopy and asked to complete a validated semiquantitative food frequency questionnaire (FFQ) [13]. In addition, subjects were asked to complete a validated physical activity questionnaire [14] to provide estimates of energy expenditure based on the frequency, duration and intensity of various types of activity (leisure-time, occupational, household, self-care), as well as an estimate of total energy expenditure from all types of activity combined. Subjects also completed the Pittsburgh Sleep Quality Index questionnaire as part of this study [15].

C. Genotyping

Eligible subjects who consented to participate in the study were required to donate a blood sample at their colonoscopy. Standard venipuncture was used to collect blood samples in tubes with EDTA as an anticoagulant. Genomic DNA was extracted from buffy coats using QIAamp DNA Blood kit (Qiagen Inc., Valencia, CA). All purified DNA samples were diluted to a constant DNA concentration in 10mmol/L Tris, 1mmol/L EDTA buffer (pH 8).

Polymorphisms in the leptin receptor gene (LEPR: rs1137100, rs1137101, rs1805096, rs6588147) and fat mass-and-obesity-associated gene (FTO: rs1421085, rs17817449, rs9939609, rs1421085, rs8044769) were detected using previously reported methods [5,6,7,10].

To ensure quality control of all genotyping results, 5% of the samples were randomly selected and genotyped by a second investigator and 1% of the samples were sequenced using a 377 ABI automated sequencer.

D. Statistical Methods

In latent variable structural equation modeling (SEM), two general sub-models are used: 1) a measurement model that develops the relationships (loadings) between the observed variables (indicators) and the latent constructs (unobserved variables); and, 2) a structural model that develops the relationships (path coefficients) between the latent variables. The general form of the measurement model is as follows [16]:

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (1)$$

where: $\mathbf{y}$ = $p \times 1$ vector of observed variables;

$\boldsymbol{\eta}$ = $m \times 1$ vector of latent random variables;

$\boldsymbol{\varepsilon}$ = $p \times 1$ vector of measurement errors for $\mathbf{y}$; and,

$\mathbf{\Lambda}_y$ = $p \times m$ matrix of coefficients relating $\mathbf{y}$ to $\boldsymbol{\eta}$.

The general form of the measurement model can be modified to include observed predictor variables (covariates) by adding them as a q -dimensional vector ($\mathbf{x}$), with a $p \times q$ matrix of regression coefficients ($\mathbf{K}$) and a p -dimensional vector of intercepts ($\mathbf{v}$):

$$\mathbf{y} = \mathbf{v} + \mathbf{\Lambda}_y \boldsymbol{\eta} + \mathbf{K}\mathbf{x} + \boldsymbol{\varepsilon} \quad (2)$$

The covariance matrix of $\boldsymbol{\varepsilon}$ is denoted $\boldsymbol{\Theta}$ and we assume $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Theta})$. The general form of the structural model imposes constraints such that [16]:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta} \quad (3)$$

where: $\boldsymbol{\eta} = m \times 1$ vector of latent random variables;

$\mathbf{B} = m \times m$ matrix of path coefficients; and,

$\boldsymbol{\zeta} = m \times 1$ vector of errors or disturbances in the endogenous (dependent) latent variables.

The structural model can also be modified to include covariate effects by adding a q -dimensional vector of covariates ($\mathbf{x}$), an $m \times q$ matrix of regression coefficients ($\boldsymbol{\Gamma}$) and an m -dimensional vector of intercepts ($\boldsymbol{\alpha}$):

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta} \quad (4)$$

The covariance matrix of $\boldsymbol{\zeta}$ is denoted $\boldsymbol{\Psi}$ and we assume $\boldsymbol{\zeta} \sim N(0, \boldsymbol{\Psi})$.

In our latent gene construct approach, we let a latent variable (e.g., η_1 , depicted as an oval in Fig. 1) represent the overall variation in a gene, which we formally describe by multiple SNPs in that gene (e.g., y_1 , y_2 , and y_3 , depicted as rectangles in Fig. 1). We describe relationships between the individual SNPs and the latent gene construct with standardized loadings (e.g., λ_{11} , λ_{12} , and λ_{13} in Fig. 1); and, describe relationships between latent gene constructs with standardized path coefficients (e.g., β_{12} in Fig. 1). We also quantify measurement error in the SNP indicators (e.g., ε_1 , ε_2 , and ε_3 in Fig. 1) and errors in endogenous latent constructs (e.g., ζ_2 in Fig. 1).

The null hypothesis (H_0) being tested is that, if the conceptualized model were correct, the population covariance matrix of the observed variables, $\boldsymbol{\Sigma}$, would be exactly reproduced by the covariance matrix as a function of the model parameters, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ (i.e., $H_0: \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$). Since the population covariance matrix is not known, covariance-based SEM aims to fit the model by minimizing the difference between the sample covariance matrix ($\mathbf{S}$) and the covariance matrix predicted by the model parameters ($\boldsymbol{\Sigma}(\boldsymbol{\theta})$) using a fitting function. Maximum likelihood (ML) estimation fitting functions (F_{ML}) of the following general form can be used for global optimization [16]:

$$F_{ML} = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \text{trace}[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}] - \log|\mathbf{S}| - (p+q) \quad (5)$$

where: $p + q$ = total number of observed variables.

Traditional ML fitting functions require the rigid distributional assumption of multivariate normality (MVN) and independence of observations. However, when $\mathbf{y}$ comprises categorical variables (e.g., genotypes), a conditional normality assumption for ($\eta|\mathbf{y}$) avoids the assumption of full multivariate normality and allows non-normality in $\boldsymbol{\eta}$; and, when all dependent observed variables ($\mathbf{y}$) are categorical, the covariance matrix can be expressed as follows [17]:

$$\boldsymbol{\Sigma}(\eta|\mathbf{y}) = \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B})'^{-1}\boldsymbol{\Lambda}_y^{-1}\boldsymbol{\Theta} \quad (6)$$

where: $\mathbf{I}$ = identity matrix; and,

$\mathbf{I} - \mathbf{B}$ = must be non-singular.

Simulation studies have shown that traditional and modified ML estimators are robust to violations of MVN and, even under conditions of severe non-normality, ML parameter estimates are still consistent [18,19]. In this work, we used the robust ML estimator (MLR) in Mplus v5.1, which has been shown to obtain consistent and efficient estimates under non-normality [20].

Prior to conducting the SEM, we evaluated the distribution of the variables and variables with extreme skew or kurtosis (as determined by visual inspection of histograms and quantile-quantile (Q-Q) plots as well as the Shapiro-Wilk's and Kolmogorov-Smirnov tests) were natural log transformed to approximate a normal distribution. To build and evaluate the measurement model (i.e., the individual latent constructs), we used a combination of factor analysis tools (eigenvalues, scree plots, factor patterns, Cronbach's alpha). For the latent gene constructs, we also utilized linkage disequilibrium (LD) information (Haploview v4.1) and retained all SNPs genotyped in a given gene to devise the construct unless the SNP created a linear dependency and provided redundant information.

To evaluate overall model fit, we use several indices, including the chi-square goodness of fit index, whereby, if the null hypothesis is correct, the minimum fitting function value ($\mathbf{F}$) multiplied by the sample size (N) converges to a χ^2 with $(s-t)$ degrees of freedom (df) [21]:

$$\chi^2_{(s-t)} = (N - 1) \mathbf{F} [\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta})] \quad (7)$$

where: s = number of non-redundant elements in $\mathbf{S}$; and,

t = total number of parameters to be estimated.

Because the chi-square test is very sensitive to sample size and MVN violations, several alternative descriptive fit indices have been developed. The root mean squared error of approximation (RMSEA) fit index represents dispersal of data to model discrepancy (or misfit of the model) across degrees of freedom and values of less than or equal to 0.06 represent good model fit [21,22]. The Comparative Fit Index (CFI) is an incremental fit index determined independent of sample size and appears to perform better than RMSEA in simulation studies [23] CFI values greater than or equal to 0.95 and 0.90 represent good and acceptable model fit, respectively [22,24]. The CFI has also been shown to have acceptable rejection rates across models at these values, including models with binary dependent variables, when the sample size is greater than or equal to 250 [25]. The standardized root mean square residual (SRMR) is an index based on fitted residuals or discrepancies between $\mathbf{S}$ and $\mathbf{\Sigma}(\boldsymbol{\theta})$. SRMR values less than or equal to 0.08 and 0.10 represent good and acceptable fit, respectively [24].

III. Results

As shown in Table 1, the study population consisted of 110 subjects with polyps and 296 without polyps. 247 subjects were Caucasian, 151 African-American and 8 were of "Other" ethnicity. There were 242 females and 164 males and their mean age was 56.0 ± 9.9 (s.d.) years. On average, the population was overweight with a mean BMI of 29.63 ± 7.02 kg/m² and waist circumference of 99.98 ± 16.96 cm. The mean total daily energy intake was 2122.39 ± 1589.00 kcal/day and intake of saturated and trans fats was 23.47 ± 20.19 g/day and 0.93 ± 0.89 g/day, respectively. Subjects reported sleeping $\sim 7.03 \pm 2.10$ hours/day and participating in leisure, recreational, household and other activities for 7.67 ± 3.23 , 0.89 ± 1.02 , 3.98 ± 2.45 and 0.22 ± 0.38 hours/wk, respectively.

In terms of the measurement model, we devised the latent gene constructs using a battery of techniques to balance parsimony and construct reliability. For example, by inspection of the

linkage disequilibrium (LD) plot shown in Fig. 2, we could have chosen any of the following three FTO SNPs in 'Block 1' to represent the variation in this 7 kb genomic area: rs17817449; rs8050136; or, rs9939609. The general redundancy in the information provided by these three SNPs in 'Block 1' can also be verified by evaluating the individual associations between each SNP on BMI and waist circumference, as depicted in Tables 2 and 3, respectively. We also note that the correlation between rs17817449 and rs8050136 was nearly 1.0 ($\rho=0.99$; $p<0.0001$), which would create a linear dependency, whereby the effects between these two SNPs could not be distinguished; thus, one of these two SNPs (rs17817449, rs8050136) must be excluded. Furthermore, although a slightly higher construct reliability can be achieved using four FTO SNPs (rs1421085, rs17817449, rs9939609, rs8044769; Cronbach's $\alpha=0.88$) versus three SNPs (rs1421085, rs9939609, rs8044769; Cronbach's $\alpha=0.78$), using three SNPs provides good coverage of the region/block, provides good reliability and is more parsimonious. Thus, we chose rs9939609 to represent 'Block 1' in our final model because it has been the most well studied FTO SNP; and, used rs1421085, rs9939609, and rs8044769 to represent the FTO gene as a latent construct in the full SEM model (discussed below).

When modeling constructs for the leptin receptor (LEPR) gene (defined by SNPs: rs1137100, rs1137101, rs1805096, rs6588147) and the fat mass-and-obesity-associated (FTO) gene (defined by SNPs: rs9939609, rs1421085, rs8044769), together with demographic (age, race, gender), physical activity, diet and sleep constructs (Fig. 3), we observed a good fitting model by several goodness-of-fit indices ($\chi^2=460.14$, $df=208$, $p<0.05$; CFI=0.92, RMSEA=0.06; SRMR=0.07). In this full model (Fig. 3), we observed an inverse association between the FTO and energy imbalance ('obesity': defined by BMI and waist circumference) constructs ($\beta_{std}=-0.13 \pm 0.06$; $p=0.03$), which was much stronger compared to that observed in a reduced model (not shown) with only the FTO and LEPR constructs and demographic variables ($\beta_{std}=-0.05 \pm 0.03$; $p=0.08$). We also observed an inverse association between the LEPR and physical activity construct, which was defined by leisure, household, recreational and occupational daily energy expenditure measures ($\beta_{std}=-0.15 \pm 0.04$; $p<0.01$); and, a marginal association between FTO and the 'bad' dietary fat construct, which was defined by dietary intake of saturated and trans fat measures ($\beta_{std}=-0.06 \pm 0.03$; $p=0.09$). Marginal associations were also observed between physical activity and energy imbalance ($\beta_{std}=-0.23 \pm 0.13$; $p=0.09$) and sleep and energy imbalance ($\beta_{std}=-0.16 \pm 0.08$; $p=0.06$) constructs. Interestingly, removing FTO revealed a marginal association between the LEPR and obesity ($\beta_{std}=0.24 \pm 0.14$; $p=0.09$), which was not present when FTO was in the model (not shown).

IV. Discussion

Our results show that by modeling the FTO and LEPR gene constructs together with diet, physical activity and sleep factors, the strength of the association between the FTO and energy imbalance ('obesity') constructs was strengthened compared to that association observed in a reduced model with only the FTO and LEPR gene constructs and demographic variables. Furthermore, we observed several significant paths directly from the gene constructs to their more proximal function rather than directly to the comprehensive phenotype of energy imbalance ('obesity') including an association between the LEPR gene and physical activity constructs and an association between the FTO gene and 'bad' dietary fat intake, which may serve as a surrogate measure for satiety. Thus, we have illustrated that our latent gene construct SEM approach to modeling biological systems such as energy balance enables a better mathematical representation of the relationships between the multiple factors involved in the system(s).

Upon removing the FTO gene construct from the model, an association was revealed between the LEPR gene and the energy imbalance ('obesity') construct, which was not present when modeling the FTO and LEPR constructs together in the same model. Interestingly, statistically significant associations were also observed when examining individual LEPR SNPs on BMI in a generalized linear model adjusted for age, race and gender. Taken together, these results illustrate that our approach provides for improved control of confounding compared to standard regression methods and enables the synthesis of single SNP information to estimate overall gene effects while capitalizing on the substantial level of correlation between SNPs in a gene.

Acknowledgments

We thank Drs. Graham Casey and Sarah Plummer at the University of Southern Carolina for performing the genotyping in this study.

This work was supported in part by NIH NCI Grants K07-CA129162 and U54-CA116867.

References

- [1]. Wright S. Correlation and causation. *J Agric Res.* 1921; 20:557–585.
- [2]. Joreskog, KG.; Sorbom. LISREL IV - A general computer program for estimation of linear structural equation systems by maximum likelihood methods. University, Department of Statistics; Uppsala Sweden: 1978.
- [3]. Joreskog KG. Basic issues in the application of LISREL. *Data communications computer data analysis.* 1981; 1:1–6.
- [4]. Nock NL, Larkin EM, Morris NJ, Li Y, Stein CM. Modeling the complex gene x environment interplay in the simulated rheumatoid arthritis GAW15 data using latent variable structural equation modeling. *BMC Proc.* 2007; S1:S118–23. [PubMed: 18466459]
- [5]. Wauters M, Mertens I, Chagnon M, Rankinen T, Considine RV, Chagnon YC, et al. Polymorphisms in the leptin receptor gene, body composition and fat distribution in overweight and obese women. *Int J Obes Relat Metab Disord.* 2001; 25(5):714–20. [PubMed: 11360155]
- [6]. Hinney A, Nguyen TT, Scherag A, Friedel S, Bronner G, et al. Genome Wide Association (GWA) Study for Early Onset Extreme Obesity Supports the Role of Fat Mass and Obesity Associated Gene (FTO) Variants. *PLoS ONE.* 2007; 2(12):e1361. [PubMed: 18159244]
- [7]. Kring SI, Holst C, Zimmermann E, Jess T, Berentzen T, Toubro S, et al. FTO gene associated fatness in relation to body fat distribution and metabolic traits throughout a broad range of fatness. *PLoS ONE.* 2008; 3(8):e2958. [PubMed: 18698412]
- [8]. Paracchini V, Pedotti P, Taioli E. Genetics of leptin and obesity: a HuGE review. *Am J Epidemiol.* 2005; 162(2):101–14. [PubMed: 15972940]
- [9]. Peters T, Ausmeier K, Rther U. Cloning of Fatso (Fto), a novel gene deleted by the Fused toes (Ft) mouse mutation. *Mamm Genome.* 1999; 10:983–6. [PubMed: 10501967]
- [10]. Wardle J, Carnell S, Haworth CM, Farooqi IS, O'Rahilly S, Plomin R. Obesity associated genetic variation in FTO is associated with diminished satiety. *J Clin Endocrinol Metab.* 2008; 93(9):3640–3. [PubMed: 18583465]
- [11]. Wardle J, Llewellyn C, Sanderson S, Plomin R. The FTO gene and measured food intake in children. *Int J Obes (Lond).* 2009; 33(1):42–5. [PubMed: 18838977]
- [12]. Moghaddam AA, Woodward M, Huxley R. Obesity and risk of colorectal cancer: a meta-analysis of 31 studies with 70,000 events. *Cancer Epidemiol Biomarkers Prev.* 2007; 16(12):2533–47. [PubMed: 18086756]
- [13]. Martinez ME, Marshall JR, Graver E, Whitacre RC, Woolf K, Ritenbaugh C, Alberts DS. Reliability and validity of a self-administered food frequency questionnaire in adenoma recurrence. *Cancer Epidemiol Biomarkers Prev.* 1999; 8(10):941–6. [PubMed: 10548325]

- [14]. Staten LK, Taren DL, Howell WH, Tobar M, Poehlman ET, Hill A, Reid PM, Ritenbaugh C. Validation of the Arizona Activity Frequency Questionnaire using doubly labeled water. *Med Sci Sports Exerc.* 2001; 33(11):1959–67. [PubMed: 11689750]
- [15]. Buysse OI, Reynolds CF III, Monk TH, Berman SR, Kupfer OJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatric Research.* 1989; 28(2): 193–213.
- [16]. Bollen, K. *Structural equations with latent variables.* Wiley; New York, NY: 1989.
- [17]. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika.* 1984; 49:115–132.
- [18]. Boomsma, A.; Hoogland, JJ. The robustness of LISREL modeling revisited. In: Cudeck, R.; du Toit, S.; Sörbom, D., editors. *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog.* Scientific Software International; Chicago: 2001. p. 139-168.
- [19]. Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling.* 2002; 4:599–620.
- [20]. Mplus v4.1 User's Guide. May. 2006 <http://www.statmodel.com/ug excerpts.shtml>
- [21]. Hancock, GR.; Mueller, RO. *Structural equation modeling: a second course.* Information Age Publishing, Inc.; Greenwich, CT:
- [22]. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling.* 1999; 6:1–55.
- [23]. Bentler PM. Comparative fit indexes in structural models. *Psychol Bull.* 1990; 107(2):238–46. [PubMed: 2320703]
- [24]. Kline, RB. *Principles and Practice of Structural Equation Modeling.* Guilford Press; New York: 2005. Measurement models and confirmatory factor analysis; p. 133-45.
- [25]. Yu, CY. Dissertation. University of California; 2002. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. <http://www.statmodel.com/papers>

Figure 1.
Latent Gene Construct Approach using the Statistical Framework of Structural Equation Modeling (see text for notation details).

Figure 2. Linkage Disequilibrium as R^2 for FTO SNPs rs1421085, rs17817449, rs8050136, rs9939609 and rs8044769.

Figure 3. Model of Energy Imbalance ('Obesity') on Colon Polyps using a Latent Gene Construct Structural Equation Modeling (SEM) Approach. Model results in good overall fit ($\chi^2=460.14$, $df=208$, $p<0.05$; CFI=0.92, RMSEA=0.06; SRMR=0.07). Standardized coefficients and their standard errors shown above arrows. Blue=genes; Green=environmental factors; Red=traits. ** $p\leq 0.05$; * $p\leq 0.10$. Residuals not shown for clarity.

TABLE I**Characteristics of The Colon Polyp Case-Control Study Population**

Characteristic	Cases (n=110)	Controls (n=296)
Age (years) ^a	58.65 ± 10.28	54.49 ± 9.45 ^b
Females	58 (52.73%)	184 (62.16%) ^b
Caucasians	54 (49.09%)	193 (65.20%)
African-Americans	52 (47.27%)	99 (33.45%) ^b
BMI (kg/m ²) ^a	29.55 ± 6.09	29.50 ± 7.19
Waist (cm) ^a	102.00 ± 17.13	98.20 ± 16.63 ^b
FTO rs1421085 ^c Intron: TT	58 (52.73%)	145 (48.99%)
TC	40 (36.36%)	118 (39.86%)
FTO rs17817449 ^c Intron: TT	41 (37.27%)	102 (34.46%)
TG	49 (44.55%)	149 (50.34%)
FTO rs8050136 ^c Intron: CC	37 (33.64%)	100 (33.78%)
CA	52 (47.27%)	149 (50.34%)
FTO rs9939609 ^c Intron: TT	34 (30.91%)	93 (31.42%)
TA	51 (46.36%)	149 (50.34%)
FTO rs8044769 ^c Intron: CC	52 (47.27%)	101 (34.12%)
CT	46 (41.82%)	146 (49.32%) ^b
LEPR rs1137100 ^c Lys 109 Arg: AA	67 (60.91%)	185 (62.50%)
AG	38 (34.55%)	95 (32.09%)
LEPR rs1137101 ^c Glu223Arg: GG	31 (28.18%)	82 (27.70%)
GA	55 (50.00%)	148 (50.00%)
LEPR rs1805096 ^c Intron: GG	37 (33.64%)	87 (29.39%)
GA	55 (50.00%)	161 (54.39%)
LEPR rs6588147 ^c Intron: AA	65 (59.09%)	155 (52.36%)
AG	39 (35.45%)	115 (38.85%)

^a Mean ± standard error of the mean.

^b Significant (p ≤ 0.05) difference between cases & controls using chi-square or t-test, as applicable.

^c Variant/Variant genotypes not shown due to space limitations but can be determined by summing the other genotypes provided and subtracting from the total no. of cases (110) or controls (296).

TABLE II

ASSOCIATIONS BETWEEN FTO SNPs AND BMI

SNP	Body Mass Index (BMI: kg/m ²) ^{a, b}		
	Wild Type/ Wild Type ^c	Wild Type/ Variant ^c	Variant / Variant ^c
rs1421085 Intron: T → C	30.74 ± 0.47	28.47 ± 0.54; p=0.002	27.96 ± 0.99; p=0.01
rs17817449 Intron: T → G	29.49 ± 0.57	29.48 ± 0.48; p=0.99	30.00 ± 0.82; p=0.61
rs8050136 Intron: C → A	29.36 ± 0.58	29.51 ± 0.48; p=0.83	29.96 ± 0.80; p=0.54
rs9939609 Intron: T → A	29.16 ± 0.60	29.69 ± 0.48; p=0.49	29.77 ± 0.53; p=0.52
rs8044769 Intron: C → T	30.48 ± 0.54	29.24 ± 0.49; p=0.09	28.14 ± 0.85; p=0.02

^a Model adjusted for age, race and gender.^b Mean BMI ± standard error (s.e.) of the mean.^c Wild Type/Wild Type= homozygous for wild type allele; Wild Type/Variant=heterozygous; Variant/Variant=homozygous for variant allele.

TABLE III

ASSOCIATIONS BETWEEN FO SNPS AND WAIST

SNP	Waist Circumference (cm) ^{a, b}		
	Wild Type/ Wild Type ^c	Wild Type/ Variant ^c	Variant/ Variant ^c
rs1421085 Intron: T → C	101.63 ± 1.15	96.95 ± 1.30; p=0.007	96.93 ± 2.46; p=0.08
rs17817449 Intron: T → G	98.90 ± 1.38	99.04 ± 1.17; p=0.94	101.10 ± 2.05; p=0.37
rs8050136 Intron: C → A	98.38 ± 1.41	99.38 ± 1.16; p=0.59	100.96 ± 1.99; p=0.29
rs9939609 Intron: T → A	97.93 ± 1.45	99.75 ± 1.17; p=0.33	100.28 ± 1.14; p=0.32
rs8044769 (R) Intron: C → T	101.97 ± 1.33	99.00 ± 1.18; p=0.02	96.61 ± 2.08; p=0.03

^a Model adjusted for age, race and gender.^b Mean BMI ± standard error (s.e.) of the mean.^c Wild Type/Wild Type= homozygous for wild type allele; Wild Type/Variant=heterozygous; Variant/Variant=homozygous for variant allele.