Modeling pleiotropy: Gee-2 linkage and association joint analysis of adolescent alcohol and cigarette consumption

Rumi Kato Price

Nathan K. Risk

Jonathan Corbett

Modeling Pleiotropy: GEE-2 Linkage and Association Joint Analysis of Adolescent Alcohol and Cigarette Consumption*

Rumi Kato Price, PhD, MPE
Nathan K. Risk, MA
Jonathan Corbett, PhD
March 2006

Abstract

Background: We previously demonstrated the utility of the generalized estimation equations covariance modeling (GEE-2) for the joint linkage & association analysis of single phenotypes for family data. **Objective**: To model jointly the association of single genetic variants and environmental and psychiatric covariates with, and the linkage of single genetic variants to, the co-occurrence of alcohol and nicotine consumption phenotypes. **Methods**: Using the National Longitudinal Study of Adolescent Health (Add Health) data, the two phenotypes are the typical number of alcoholic drinks consumed in the past year, and the typical number of cigarettes smoked per day in the past month. DAT1, DRD4, 5HTT, CYP2A6, DRD2 are successively included in the GEE-2 models with covariates. **Results**: Good evidence of allelic transmission was observed for DAT1, CYP2A6 and DRD2 using the combined alcohol and cigarette consumption phenotypes. However, there was no evidence of marker-specific association with either single or multiple phenotypes. **Summary**: GEE-2 linkage & association joint analysis is a flexible method to examine pleiotropic linkage and association for genetic markers, controlling for environmental and other covariates.

Introduction

<u>Pleiotropy</u>: Originally, it meant a single gene is responsible for a number of distinct and seemingly unrelated phenotypic effects. For complex disorders, multiple genes can be assumed to influence multiple phenotypes (e.g., obesity, gallbladder disease, and non-insulin-dependent diabetes mellitus among New World Native Americans (Weiss, 1993)). Substance use/abuse phenotypes are likely to be influenced by multiple overlapping genes.

<u>Linkage & association joint analysis</u>: Using a traditional association analysis approach involving cases and controls: stratification can lead to an erroneous conclusion of association (i.e., both disease frequencies and allele frequencies are different among subpopulations). Alternatively, linkage and association joint analysis uses a family-based design where the null hypothesis of no association is rejected only in the presence of both association and linkage. TDT (transmission-disequilibrium test) to QTDT (quantitative TDT) are among the most commonly used models. GEE-2 has been proposed as a unified approach applicable to both family and population data. QTDT and GEE-2 provide similar results in simulated data (Price et al, under review)

Methods (1): Data Source
National Longitudinal Study of Adolescent Health (Add Health)

- **Objectives (Resnick, et al., 1997):** Primary focus on influences of social environments on adolescent health.
- **Wave 1 Methods:** In-School (Add Health S95) survey N=90,118; In-Home (Add Health-H95) survey N=20,105. Students were in grades 7-12. The surveys were completed in 1994-5 (Figure 1).
- **Wave 2 methods:** In-Home W2 N=14,738 included a subset sample of W1 In-Home sample. The survey completed in 1996.
- **Wave 3 methods:** In-Home W3 N=15,197 included a subset sample of W1 In-Home sample. Data collection in 2001-2. DNA typing performed on the genetic subsample including twins, full-sibs, half-sibs and others (n=2,574) (Figure 1). Full and half-sibling statuses were based on self-reports at W1, and 11 microsatellites collected at W1 and W3 were used to determine twin zygosity. Sibling clustering is adjusted in our analyses. The sample is not weightable due to inclusion of respondents from outside the original sampling frame.

Methods (2): Add Health - Six Markers (Figure 2)

- **Dopamine transporter (DAT1, locus SLC6A3):** VNTR includes 3-11 copies. A9 and A10 (440, 480 bp) are most common accounting for >90%. VNTR affects translation of the DAT protein in human striatum.

- **Dopamine D4 receptor (DRD4):** VNTR includes 2-11 copies. A4 and A7 are most common. VNTR results in a variation in the 3rd cytoplasmic loop in the receptor protein.

- **Serotonin transporter (5HTT, locus SLC6A4):** Long allele (L, 528bp) is thought to produce 3 times basal activity compared to short allele (S, 484bp) of protein production.

- **Moamine Oxdase A promotor (MAOA):** VNTR 3 and 4 repeats account for 95% in males. The gene product is responsible for degradation of dopamine, serotonin and norepinephrine.

- **Cytochrome P450 2A6 (CYP2A6*2):** SNP *2 nonsynonymous T->A in codon 60 (exon 3) results in substitution of leucine for histidine, which produces catalytically inactive protein. Low gene frequencies.

- **Dopamine D2 receptor (DRD2 TaqIA):** A1 (304bp) is less common than A2(178p). A1 contains a point mutation C->T, which eliminates the TaqI site.

- Evidence of association with alcohol, nicotine, illicit drug, _or_ aggression phenotypes from 70+ articles.

- MAOA omitted in subsequent analyses because it is sex-linked.

Methods (3) Phenotypes and Covariates

Alcohol consumption phenotype (n=1,287):
- "Have you had a drink of beer, wine or liquor –not just a sip or a taste of someone else's drink –more than 2 or 3 times in your life?" Total retained in sample (yes)=1,363; Others (no)=1,180 → skip out.
- "Think of all the times you have had a drink during the past 12 months. How many drinks did you have each time?" If R drank in his or her lifetime but not in the past year → 0. If R claimed to have had 19 or more drinks in a usual time → missing.

Cigarette consumption phenotype (n=1,012):
- "Have you ever tried cigarettes, even just 1 or 2 puffs? (No → skip out). "How old where you when you smoked a whole cigarette, for the first time?" ("Never smoked a whole cigarette" → skip out). Total retained in sample = 1,034; Others = 1,509 → missing.
- "During the past 30 days, how many cigarettes did you smoke each day on the days you smoked?" If R claimed to have smoked 60 or more cigarettes a day → missing.

Covariates:
- Gender, age, race (Caucasian, African American, Asian and Other), conduct symptoms, college aspiration, friends who use alcohol, who smoke cigarettes, and being sexual active were chosen after the first-stage means-only analyses.

Methods (4) GEE-2

- GEE-2 means equation for individuals (Thomas et al., 1999):

$$\mu_{ij} = \beta_0 + Z'_{ij}\beta$$

- GEE-2 covariance equation for families:

$$C_{ijk} = (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik}) = \alpha_0 + \pi_{ijk}\alpha_1 + x_{ijk}\alpha_2 + \varepsilon$$

where $i = 1 \ldots M$ nuclear families with n_i children, $j,k = 1 \ldots n_i$ children, Z_{ij} = covariates for child j in family i, β = estimated parameters for means, C_{ijk} = cross product of jth and kth child's phenotypic value, π_{2ijk} = proportion of alleles IBD, x_{ijk} = pair specific covariate, α_x = estimated parameters for covariance.

- The two equations are solved with the estimating equations:

$$\sum_i [Y_i - \mu_i(\beta)]'W_i^{-1}\frac{\partial \mu_i}{\partial \beta} = 0 \quad \sum_i [C_i - \Sigma_i(\alpha)]'V_i^{-1}\frac{\partial \Sigma_i}{\partial \alpha} = 0$$

where $\Sigma_i(\alpha) = E(C_i \mid \alpha)$

- Advantages of GEE- 2: It can be applied to both binary and quantitative phenotypes, and is useful for modeling pleiotropy (multiple correlated phenotypes) and longitudinal phenotypes. Also it is possible to include a variety of measures in the covariance model (e.g. kinship coefficients). Modeling can be staged (Figure 3). SAS macro can automate iterations between the means and covariance equation estimations at a later stage.
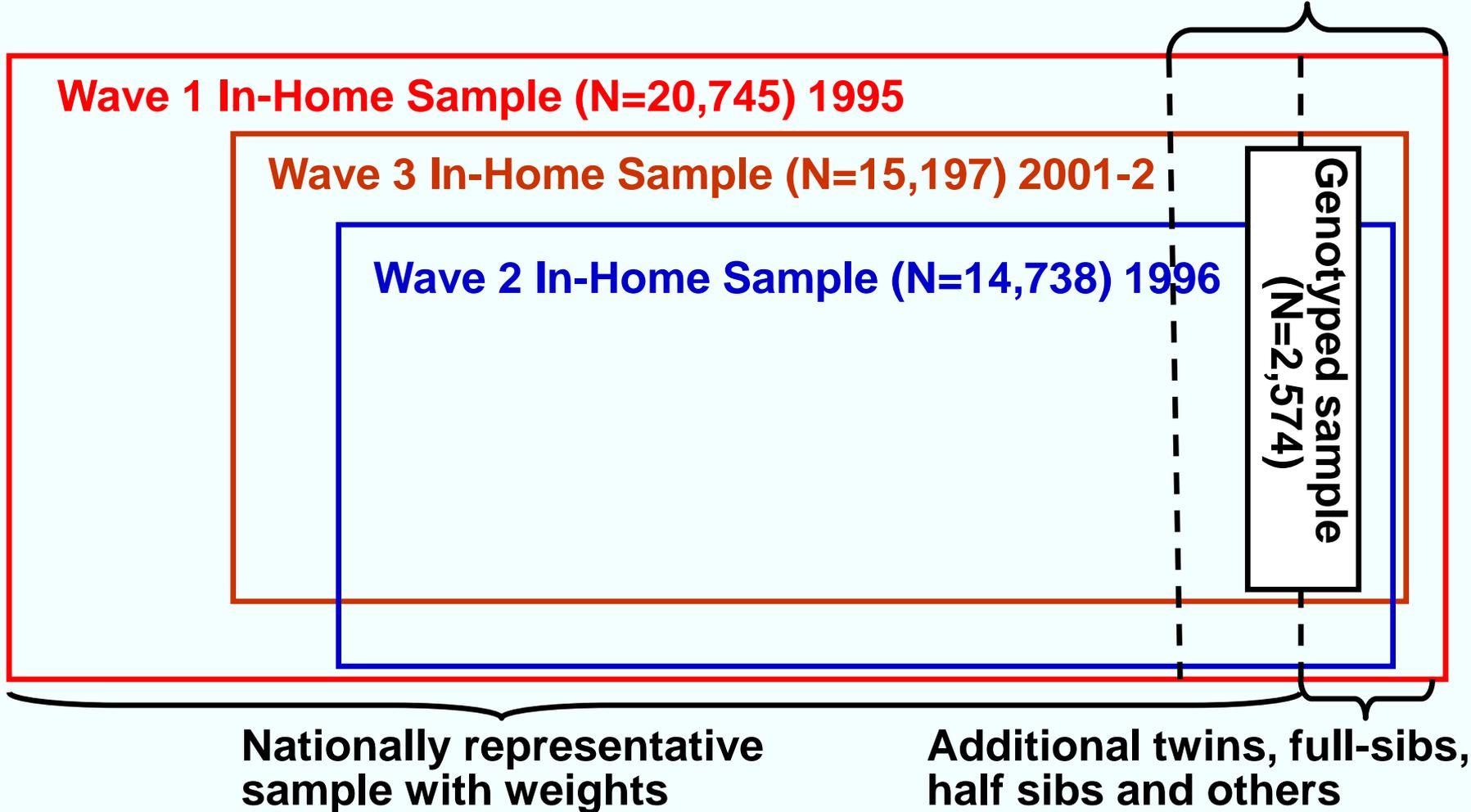
Results

Send an email request to price@rkp.wustl.edu

Extended References

Cardon LR, Abecasis GR. Some properties of a variance components model for fine-mapping quantitative trait loci. Behav Gen 2000; 30: 235-43.

Chen WM, Broman KW, Liang KY. Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genet Epidemiol 2004; 26: 265-72.

Diggle P, Liang K-Y, Zeger S. Analysis of Longitudinal Data. Oxford, UK: Oxford University Press, 1994.

Fulker DW, Cherny SS, Sham PC, Hewitt JK. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 1999; 64: 259-67.

Hopfer CJ, Timberlake D, Haberstick B, Lessem JM, Ehringer MA, Smolen A, Hewitt JK. Genetic influences on quantity of alcohol consumed by adolescents and young adults. Drug Alcohol Depend 2004; 78: 187-93.

National Longitudinal Study of Adolescent Health. Research design, 1998. Available from URL:http://www.cpc.unc.edu/projects/addhealth/design.

Resnick MD, Bearman PS, Blum RW, Bauman KE, Harris KM, Jones J, Tabor J, Beuhring T, Sieving RE, Shew M, Ireland M, Bearinger LH, Urdy J. Protecting adolescents from harm: findings from the National Longitudinal Study of Adolescent Health. J Am Med Assoc 1997; 278: 823-32.

Sham PC, Purcell S. Equivalence between Haseman-Elston and variance-components linkage analyses of sib pairs. Am J Hum Genet 1001; 68: 1527-32.

Thomas DC. Statistical Methods in Genetic Epidemiology. New York, NY: Oxford University Press, 2004.

Thomas DC, Qian D, Gauderman WJ, Siegmund K, Morrison JL. A generalized estimating equations approach to linkage analysis in sibships in relation to multiple markers and exposure factors. Genet Epidemiol 1999; 17: Suppl 1; S737-42.

Weiss K, Weiss C. Genetic Variation and Human Disease: Principles and Evolutionary Approaches. Cambridge, NY: Cambridge University Press, 1995.

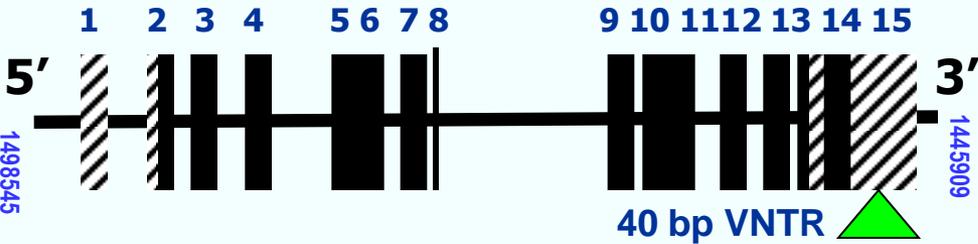# Figure 1. Add Health Sampling Scheme across 3 Waves

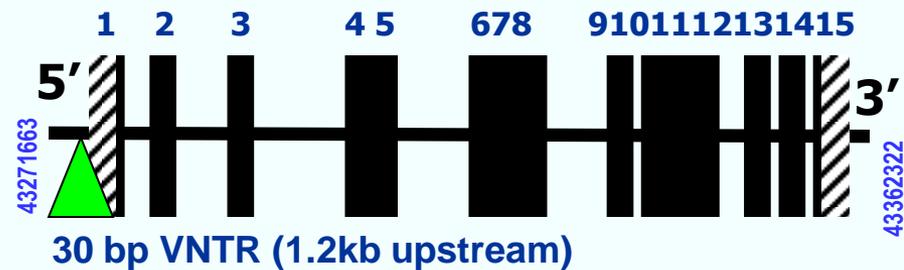**Genetically informative sample of twins, full-sibs, half sibs and others**

Wave 1 In-Home Sample (N=20,745) 1995

Wave 3 In-Home Sample (N=15,197) 2001-2

Wave 2 In-Home Sample (N=14,738) 1996

Genotyped sample (N=2,574)

Nationally representative sample with weights

Additional twins, full-sibs, half sibs and others

Figure 2. Add Health- W, Twins & Other Sibs: Six Gene Variants

DAT1 (Chr 5, 52.69kb)

MAOA (Chr X, 90.66kb)
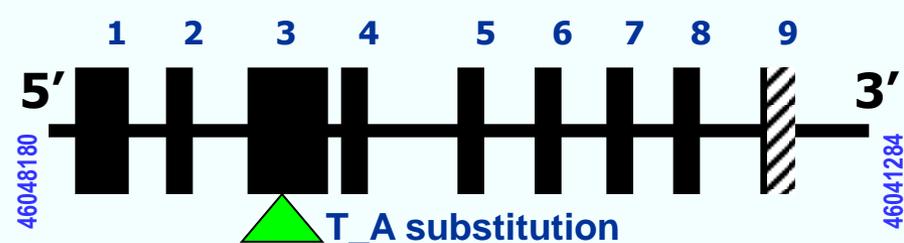30 bp VNTR (1.2kb upstream)

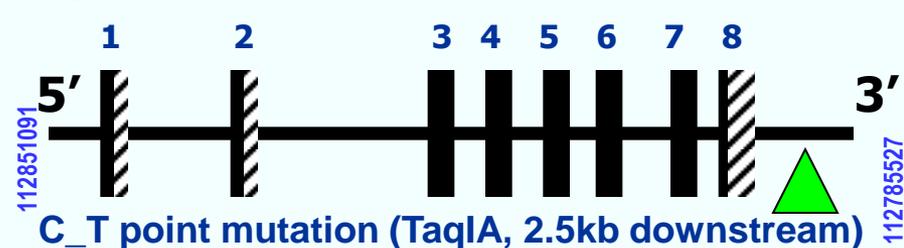40 bp VNTR

DRD4 (Chr 11, 3.40kb)
48 bp VNTR

CYP2A6 (Chr 19, 6.90kb)
T_A substitution

5HTT (Chr 17, 37.8kb)
44 bp VNTR (1.2kb upstream)
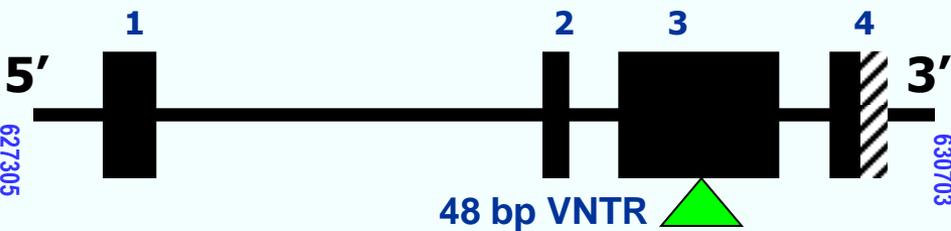
DRD2 (Chr 11, 65.58kb)
C_T point mutation (TaqIA, 2.5kb downstream)

■ Coding region   ▨ Untranslated region

Figure 3. Steps in GEE-2 Modeling

The entire genotyped samples (with singletons) are used into the initial means equation (N ≈ 1300) without the marker in the means equation. Covariances are estimated as a nuisance parameter.

The adjusted means are passed to the covariance equation (N ≈ 400 pairs) The covariances are modeled for linkage using IBD scores and sibling effects (without singletons).

Evidence of linkage!

Linkage Only

The entire genotyped samples (with singletons) are used into the means equation (N ≈ 1300) with the marker in the means equation to estimate association of the marker. Covariances are estimated as a nuisance parameter.

Evidence of association should reduce the linkage value!

The adjusted means are passed to the covariance equation (N ≈ 400 pairs) The covariances are modeled for linkage using IBD scores and sibling effects (without singletons).

Linkage and Association