Twin Research and Human Genetics

"Genome-wide meta-analysis of longitudinal alcohol consumption across youth and early adulthood"

Daniel E. Adkins, PhD, Shaunna L. Clark, PhD, William E. Copeland, PhD, Martin Kennedy, PhD, Kevin Conway, PhD, Adrian Angold, MRCPsych, Hermine Maes, PhD, Youfang Liu, PhD, Gaurav Kumar PhD, Alaattin Erkanli, PhD, Ashwin A. Patkar, MD, Judy Silberg, PhD, Tyson H. Brown, PhD, David M. Fergusson, PhD, L. John Horwood, MSc, Lindon Eaves, PhD, Edwin J.C.G. van den Oord, PhD, Patrick F. Sullivan, MD FRANZCP, E. J. Costello, PhD

CONTENTS

1. Estimating longitudinal measures of alcohol consumption

To develop our primary longitudinal measure of alcohol consumption, we employed a previously validated mixed model approach conceptualizing individual differences in consumption trajectories as random effects[1-2]. Specifically, we used a model fitting procedure to determine the optimal functional form of longitudinal, alcohol trajectories[3-4]. After identifying the optimal functional form, we estimated linear mixed models predicting alcohol consumption with fixed and random effects for age. Separate model fitting procedures were conducted for each of the 3 datasets examined (i.e., Great Smoky Mountain Study (GSMS)[5], Virginia Twin Study on Adolescent Behavioral Development (VTSABD)[6], Christchurch Health and Development Study (CHDS)[7]).

Linear mixed models are a generalization of linear regression allowing for the inclusion of individual-level random deviations (effects) other than those associated with the overall residual term. In matrix notation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \qquad \text{Eq. 1}$$

where $\mathbf{y}$ is the $n \times 1$ vector of responses, $\mathbf{X}$ is a $n \times p$ design/covariate matrix for the fixed effect $\boldsymbol{\beta}$, and $\mathbf{Z}$ is the $n \times q$ design/covariate matrix for the random effects $\mathbf{u}$. The $n \times 1$ vector of residuals $\boldsymbol{\varepsilon}$, is assumed to be multivariate normal with mean zero and variance matrix $\sigma_e^2 \mathbf{I}_n$.

The fixed portion, **Xβ**, is equivalent to the linear predictor of OLS regression. For the random portion, **Zu** + **ε**, it is assumed that the **u** has variance-covariance matrix **G** and that **u** is orthogonal to **ε** so that

$$\text{Var}\begin{bmatrix}\mathbf{u}\\ \varepsilon\end{bmatrix}=\begin{bmatrix}\mathbf{G} & 0\\ 0 & \sigma_e^2\mathbf{I}_n\end{bmatrix} \quad \text{Eq. 2}$$

The random effects **u** are not directly estimated (although, as described below, they are predicted), but instead are characterized by the elements of **G**, known as the variance components, that are estimated along with the residual variance σ_e^2. Considering **Zu** + **ε** the combined error, we see that **y** is multivariate normal with mean **Xβ** and $n \times n$ variance-covariance matrix

$$\mathbf{V} = \mathbf{ZGZ'} + \sigma_e^2\mathbf{I}_n \quad \text{Eq. 3}$$

Prior to beginning model fitting procedure, data was harmonized between studies. All three considered datasets analyzed a measure of average number of drinks in a given time period. For VTSABD, the time period was the last three months, and for GSMS and CHDS it was average drinks per week. To make models more easily comparable the consumption measure for VTSABD was divided by 13 (the number of weeks in the average 3 month period). Heterogeneity in the consumption definitions examined was not problematic, as the goal of the phenotype modeling was to estimate longitudinal variation across subjects within each study. In CHDS, drinks per weeks was coded in intervals which were converted to interval means to approximate the continuous measures assessed in VTSABD and GSMS. Also, the earliest age observed varies across studies, with the VTSABD and GSMS beginning collection of alcohol

consumption measures in childhood (ages 8-9), while the CHDS did not begin until adolescence (age 14). This study age range heterogeneity had little impact on the results, primarily serving to increase the intercept estimate in the CHDS. Sensitivity analyses truncating the age range in the VTSABD and GSMS to correspond to the CHDS yielded highly correlated slope estimates ($r>0.95$), supporting the robustness of the slope measures. Reported alcohol consumption was markedly higher among the New Zealand-based CHDS. This is likely due to the younger legal drinking age in New Zealand (age 18) versus the US (age 21) [8], in combination with the older age range considered for the CHDS. The phenotype data analyzed are summarized below in Table 1. To optimize precision in the phenotype modeling, all available phenotype information was analyzed, not only that of the genotyped subsets of subjects. Thus, the number of subjects and observation reported for the genome-wide association testing is a subset of the data analyzed to calculate longitudinal phenotypic measures.

In the current analysis, our model fitting procedure is designed to identify the optimal piecewise functional form of longitudinal alcohol consumption. This modeling assumes that the general trajectory of longitudinal alcohol consumption changes until a given age and then stabilizes. The exact age at which consumption plateaus is determined empirically as described below. To elaborate, this method does not assume an increasing or decreasing trajectory for any given subject; this is determined by the individual subject's repeated assessments. Rather, the method simply estimates a subject-specific slope capturing linear change until the estimated plateau age, after which time the slope is modeled as flat (i.e., a piecewise linear function). Thus, for

subjects that never initiate alcohol consumption, both the intercept estimate and slope estimate are equal to zero. Only the slope estimate, and not the intercept estimate, is considered an outcome in the GWAS described in the main text. This is because the intercept occurs at sufficiently young ages that its mean (fixed) effect is close to zero and there was no significant subject-specific (random intercept) variation around the mean (i.e., there is virtually no variance in alcohol consumption at the earliest ages observed).

Table 1: Descriptive statistics of alcohol consumption repeated measures

	Ages	N	Obs	Mean	SD	Range
Developmental trajectory						
GSMS	9-29	1415	9437	1.31	6.33	0-160
VTSABD	8-32	2588	5837	1.01	3.64	0-69
CHDS	14-30	1265	6921	9.11	17.20	0-60
Mean consumption						
GSMS	12-21	1345	6670	1.22	5.79	0-160
VTSABD	12-21	2337	3724	0.91	1.70	0-35
CHDS	14-21	1072	4931	6.30	14.33	0-60

Our model fitting procedure estimated the age at which alcohol consumption becomes stable. This is vital to developing high-signal measures of alcohol consumption, because to the extent the assumed trajectory functional form deviates from the actual one, s and the power to detect associations in the GWAS is proportionately diminished. For this reason, we estimate a series of models in which the assumed age at which alcohol consumption plateaus varies systematically, beginning with a model that assumes consumption stabilizes at the second earliest observed age (age 9, 10, and 15

for VTSABD, GSMS and CHDS, respectively) and then increasing age at stabilization in 1 year increments, until the maximum age in the study (age 32, 29 and 30 for VTSABD, GSMS and CHDS, respectively).

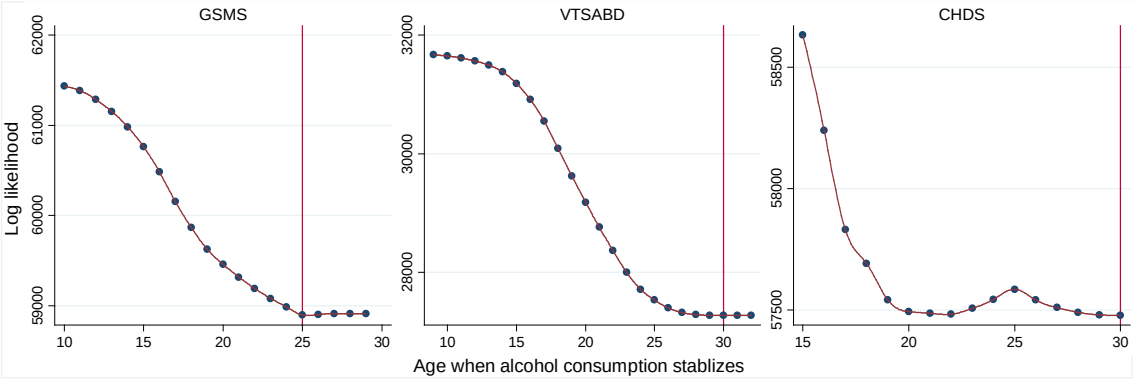
This procedure is achieved using the equation:

$$y_{ij} = \beta_0 + \beta_1 A_{ij} + u_{0i} + u_{1i} A_{ij} + e_{ij} \quad \text{Eq. 4}$$

where *i* and *j* denote the subject and assessment levels, respectively; *y* is alcohol consumption for subject *i* at assessment *j*; β_0 is the overall sample intercept; β_1 is the sample mean slope (i.e., fixed effect) of alcohol consumption; u_{0i} is the subject-specific deviation (i.e., random effect) from that overall sample intercept; u_{1i} is the subject-specific deviation from the mean alcohol consumption slope coefficient; e_{ij} is the residual for subject *i* at assessment *j*. Most importantly, A is the age variable. A is recoded in each model of the series to specify a different number of years of linear change until stable alcohol consumption (i.e., a plateau) is achieved. Thus, for the first model in the series, which assumes consumption stabilizes at the second earliest observed age (recoded as age 1), A is coded 0 at the earliest age observed in the dataset, 1 at second earliest age observed, and 1 each age thereafter. In the second model which assumes consumption stabilizes at the third earliest age observed, A is coded 0 at the earliest age, 1 at the second earliest age, and 2 for the third earliest age and remains 2 at each age thereafter. This process of incrementally increasing the number of years until the plateau is assumed continues until the oldest age observed in the dataset, in which case the model specifies constant linear change (no plateau). For each dataset, indices of model fit (log likelihoods) were collected for each model in the

A series, and then these values were minimized to determine the optimal response functional form. Figure 1 presented the graphed log likelihood values for each dataset with the optimal plateau values marked with red vertical lines. This analysis indicated

Figure 1. Identifying optimal plateau function forms for alcohol consumption trajectories



that alcohol consumption stabilized at age 30 in CHDS and GSMS, and age 25 in VTSABD. Parameter estimates for the final, preferred piecewise trajectory model are presented in Table 2.

After determining the proper functional form of the over-time alcohol consumption trajectories, we then output the random effects for trajectory measures for each of the 3 datasets. These measures quantify the deviation of each subject's alcohol consumption slope from the overall sample mean change and thus, serve as our developmental trajectory measures in the subsequent GWAS. In each model, the covariance structure of the 2 random effects was modeled as independent:

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N(0,\mathbf{G}) \qquad \text{with } \mathbf{G} = \begin{bmatrix} \sigma^2_{u0} & \\ 0 & \sigma^2_{u1} \end{bmatrix} \qquad \text{Eq.5}$$

Thus, the random parameters are multivariate normal distributed with means of zero and variance-covariance matrix **G**. The variances of the parameters are on the diagonal and the covariances, constrained equal to zero, are in the off-diagonal cells of **G**. The residual is assumed to be normally distributed with a mean of zero and variance of σ^2_e.

Table 2. Parameter estimates for final piecewise linear mixed trajectory models

	GSMS		VTSABD		CHDS	
	b	se	b	se	b	se
Age slope	0.343***	0.017	0.328***	0.01	1.154***	0.043
Intercept	-1.014***	0.092	-1.415***	0.056	2.298***	0.254
Random Age SD	0.449***	0.011	0.269***	0.005	0.914*	0.032
Random Intercept SD	0.000	0.000	0.000	0.000	0.000	0.000
Residual SD	4.922***	0.039	2.022***	0.022	14.265***	0.133
N	9437		5837		6921	
-2 Log likelihood	58895.8		27717.4		57476.8	

* p<0.05, ** p<0.01, *** p<0.001

Because random effects are not directly estimated by the mixed model, they must be predicted in an additional post-estimation step [9]. Best linear unbiased predictors (BLUPs) of the random effects **u** were obtained as:

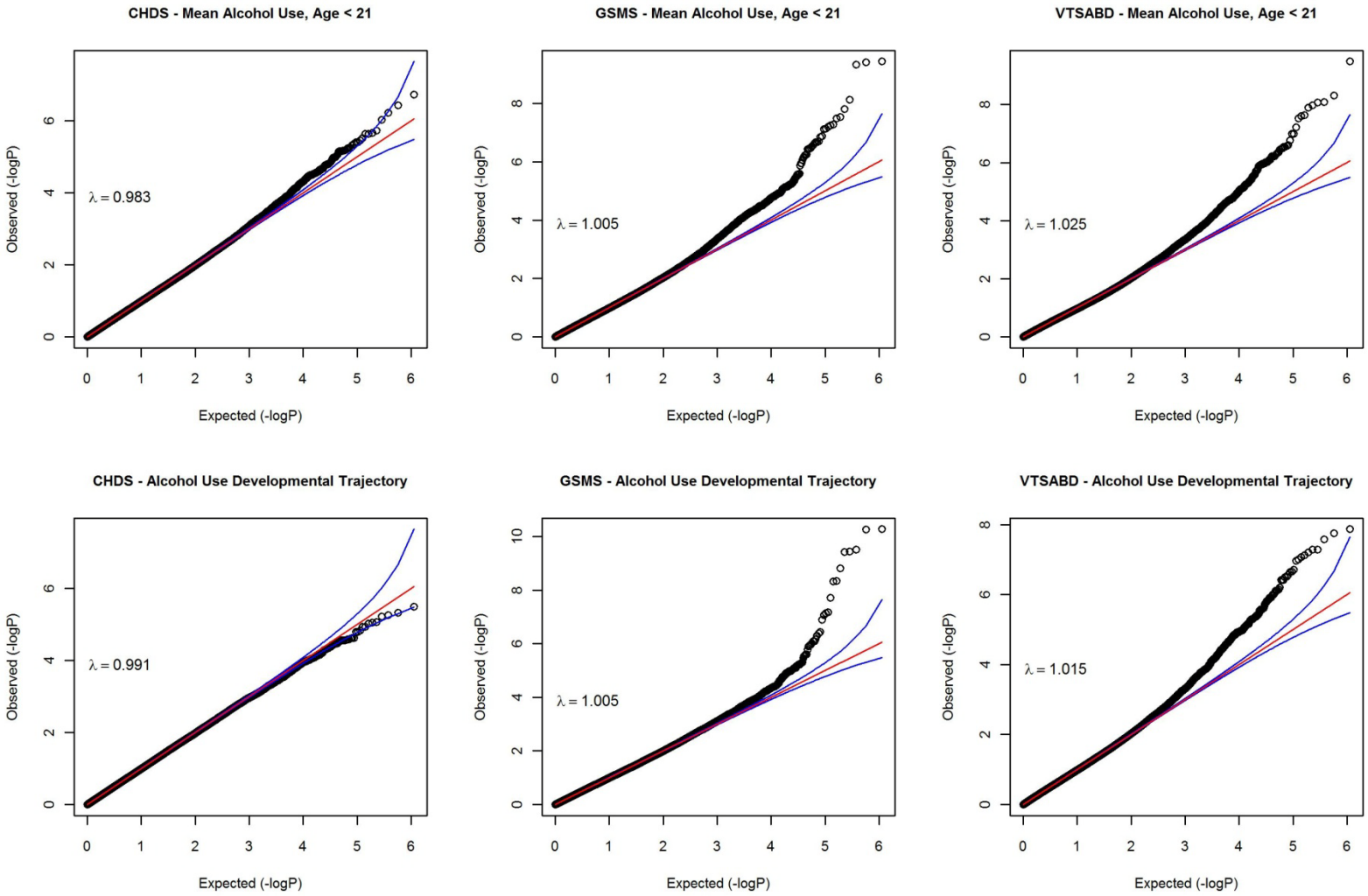
$$\tilde{\mathbf{u}} = \tilde{\mathbf{G}}\mathbf{Z}'\tilde{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \qquad \text{Eq.7}$$

where $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{V}}$ are **G** and **V** with estimates of the variance components plugged in. The EM algorithm was used for maximum likelihood estimation[10].

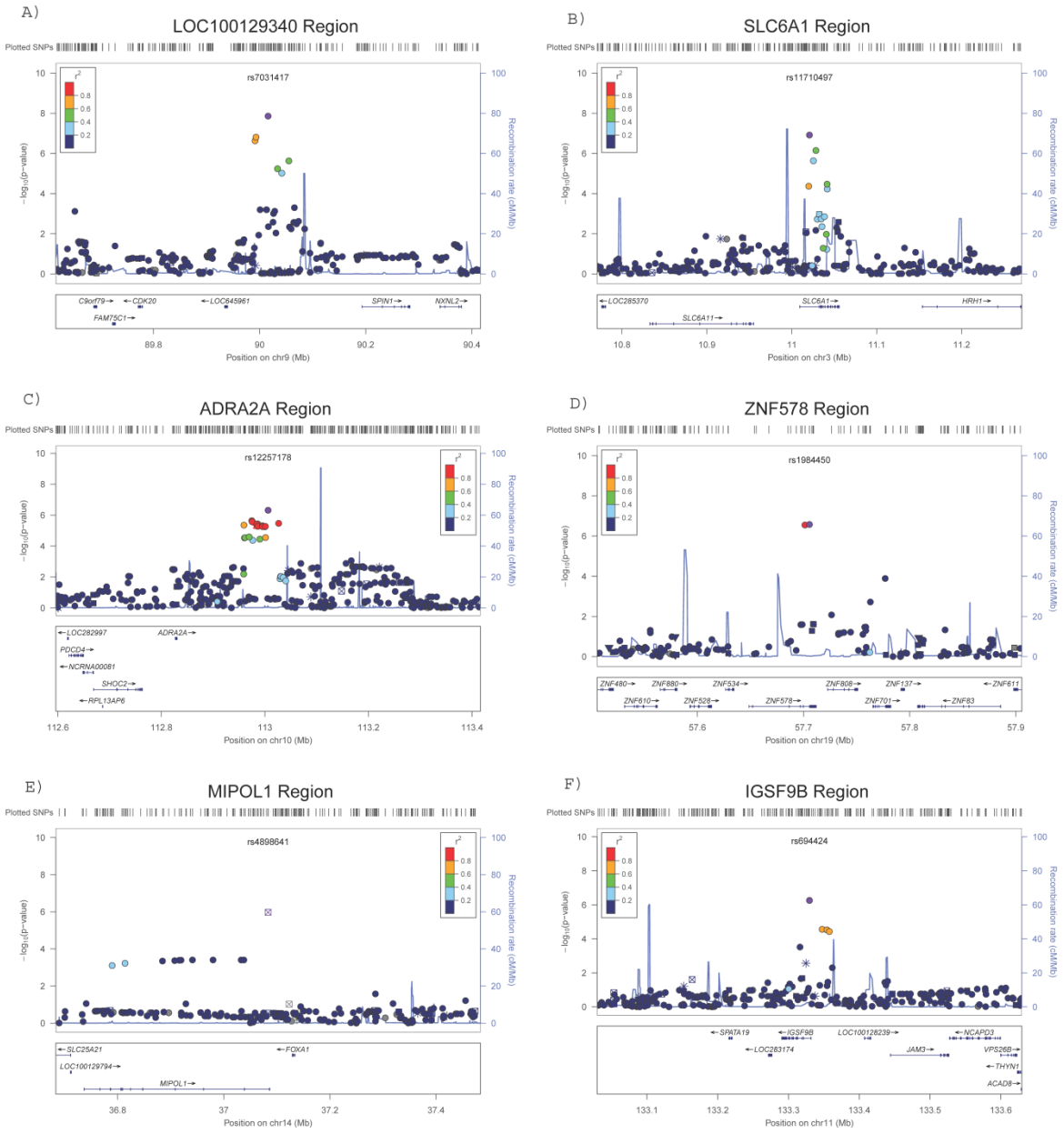
Finally, in addition to considering subject-level alcohol trajectory as an outcome in the GWAS reported in the main text, we also considered an alternative outcome, mean alcohol consumption across adolescence and the transition to adulthood (~ages 13-21). We considered this second phenotype for several reasons. First, the mean

consumption outcome provided a simpler summary of individuals' drinking behavior, and thus, greater continuity to existing literature[11-12]. Second, the mean adolescent consumption measure also has the benefit of focusing solely on a developmental period of non-normative drinking; a period which is associated with increased risk of concurrent comorbid psychiatric disorders and future substance abuse and dependence[13]. The measure was calculated simply as the mean of all alcohol consumption assessments collected in the age range. There was slight age range heterogeneity across datasets, with VTSABD and GSMS providing data for the full 12-21 age range, while CHDS began data collection alcohol consumption slightly later, leading to a 14-21 age range for this dataset. Summary information on the repeated assessments included in these measures is given in Table 1.

2. Study specific Q-Q plots



3. Regional plots for significant (q < 0.1) and suggestive(q < 0.2) associations at: A) LOC100129340; B) SLC6A1; C) ADRA2A; D) ZNF57; E) MIPOL1; F) IGSF9B.



References

1. Goldstein H. *Multilevel statistical models*. London: Arnold; 1995.

2. Searle SR, Casella G, McCulloch CE. *Variance components*: New York : Wiley, c1992.

3. van den Oord E, Adkins DE, McClay J, Lieberman J, Sullivan PF. A systematic method for estimating individual responses to treatment with antipsychotics in CATIE. *Schizophrenia Research.* Jan 2009;107(1):13-21.

4. Adkins DE, Aberg K, McClay JL, et al. Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs. *Mol Psychiatry.* Mar 2011;16(3):321-332.

5. Costello EJ, Angold A, Burns BJ, et al. The Great Smoky Mountains Study of youth - Goals, design, methods, and the prevalence of DSM-III-R disorders. *Archives of General Psychiatry.* Dec 1996;53(12):1129-1136.

6. Simonoff E, Pickles A, Meyer JM, et al. The Virginia Twin Study of Adolescent behavioral development - Influences of age, sex, and impairment on rates of disorder. *Archives of General Psychiatry.* Sep 1997;54(9):801-808.

7. Fergusson DM, Horwood LJ. The Christchurch Health and Development Study: review of findings on child and adolescent mental health. *Australian and New Zealand Journal of Psychiatry.* Jun 2001;35(3):287-296.

8. Kypri K, Voas RB, Langley JD, et al. Minimum purchasing age for alcohol and traffic crash injuries among 15-to 19-year-olds in New Zealand. *American Journal of Public Health.* Jan 2006;96(1):126-131.

9. Robinson GK. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science.* 1991;6(1):15-32.

10. Pinheiro JC, Bates DM. *Mixed-effects models in S and S-plus.* New York, NY: Springer; 2000.

11. Agrawal A, Freedman ND, Cheng YC, et al. Measuring alcohol consumption for genomic meta-analyses of alcohol intake: opportunities and challenges. *American Journal of Clinical Nutrition.* Mar 2012;95(3):539-547.

12. Grant JD, Agrawal A, Bucholz KK, et al. Alcohol Consumption Indices of Genetic Risk for Alcohol Dependence. *Biological Psychiatry.* Oct 2009;66(8):795-800.

13. Rutter M, Silberg J, O'Connor T, Simonoff E. Genetics and child psychiatry: II - Empirical research findings. *Journal of Child Psychology and Psychiatry.* Jan 1999;40(1):19-55.