SUPPLEMENTAL MATERIAL FOR PAPER:

"GENOMEWIDE ASSOCIATION FOLLOWED BY REPLICATION IMPLICATES A NOVEL GENE FOR NEUROTICISM"

CONTENT

1. CALCULATION OF THE FDR AND Q-VALUES

FDRs can be estimated in multiple ways and many standard computer packages (e.g. R, SAS) have such estimation procedures implemented. The first approach is to estimate the FDR for a chosen threshold P-value *t*. If the *m* P-values are denoted p_i, $i = 1 \ldots m$, this can be done using the formula:

$$\widehat{FDR}(t) = \frac{m\ t}{\#\{p_i \le t\}}$$

Thus, the FDR is estimated by dividing the estimated number of false discoveries (is number of tests times the probability *t* of rejecting a marker without effect) by the total number of significant markers (i.e. total number of P-values smaller than *t*) that includes the false and true positives. To avoid arbitrary choices, each of the observed P-values can be used as a threshold P-value *t*. The resulting FDR statistics are then called q-values [1;2].

Reference List

(1) Storey J, Tibshirani R. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* 2003;100:9440-9445.

(2) Storey J. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* 2003;31:2013-2035.

2. CALCULATING GENDER AND AGE MEASUREMENT INVARIANT FACTOR SCORES

Our method for calculating gender and age measurement invariant factor scores followed the general methodological approach as discussed in Neale et al. 2006. This approach involves four steps: 1) specify a factor model where individual neuroticism items are indicators of a latent/unmeasured neuroticism factor, 2) account for "real" effects of gender, age, sex, and gender by age interaction on the mean and variance of the neuroticism factor, 3) test whether there are additional effects of gender, age, sex, and gender by age interaction on the item parameters (factor loading plus thresholds), 4) estimate factor scores using a model that includes all significant effects.

Below if a more detailed description of the separate steps.

Ad (2) First, a baseline model was established without any covariate effects. This baseline model was the most parsimonious model which does not include any gender, age, sex, and gender by age interaction effects. Next, models that sequentially added gender, age, sex, and gender by age interaction in the factor mean and variance were fitted. To test whether these gender, age, sex, and gender by age interaction were significant we calculated minus two times the log likelihood differences between baseline and covariate models, which asymptotically has a chi-square distribution.

Ad (3) Specific effects on factor loadings and thresholds/means were tested using the model derived in step (2) as the baseline model.

Ad (4) Factor scores were estimated after fixing the item level parameter estimates derived from the final in step (3).

All models were fitted with the Mx package (Neale et al., 2004) using the marginal maximum likelihood (Bock & Aitkin, 1981). The latent factor distribution was specified using a ten-point Gaussian quadrature. This approach has been shown to have some distinct advantages for hypotheses testing in latent variable models (Schmitt et al., 2006).

Bock, R. D. & Aitkin, M. (1981). Marginal maximum-likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Neale, M. C., Aggen, S. H., Maes, H. H., Kubarych, T. S., & Schmitt, J. E. (2006). Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors, 31,* 1010-1034.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2004). *Mx: Statistical Modeling.* (6th ed.) Box 980126, Richmond, VA 23298: Department of Psychiatry, Medical College of Virginia, Virginia Commonwealth University.

Schmitt, J. E., Mehta, P. D., Aggen, S. H., Kubarych, T. S., & Neale, M. C. (2006). Semi-nonparametric methods for detecting latent non-normality: A fusion of latent trait and ordered latent class modeling. *Multivariate Behavioral Research, 41,* 427-443.

3. COMBINED ANALYSES AND ESTIMATION OVERALL Q-VALUES

To determine replication status in Table 3, we required the same SNP, same direction of effects, and P-values smaller than 0.05. However, this procedure is not optimal as there will be a loss of power because the stage 1 GWAS data are ignored. Furthermore, although the number of markers tested in the replication study is arbitrary, it would directly effect the number of "replications" we will obtain. In Table 4 we therefore also presented an alternative approach that has better power to detect effects and results in a statisically threshold for declaring what constitutes a replication that is more easy to intepret. Note that by combining the data automatically includes the "direction of effects" criterion as effect in opposite directs will tend to cancle out when data are combined. This alternative appraoch essentially generalizes Storey's *q*-value to the situation where data are combined across the two stages:

$$\hat{q}(t,c_1) = \frac{m\hat{p_0}\Pr\left(T_{(1+2)} > t, T_1 > c_1 \mid H_0\right)}{\#\left(\text{rejected}\right)}.$$

where c_1 is the first stage critical value, t is the second stage (combined) test statistic value, and m the number of markers. To calculate the denominator we can simply count the number of markers rejected in the replication study #(rejected) = #{ $(T_{(1+2)}, T_1) : T_{(1+2)} > t$, $T_1 > c_1$) }. For p_0 we can either use an estimate or conservatively assume it is 1, which made very little difference in this case. The main challenge is the calculation of Pr($T_{(1+2)} > t$, $T_1 > c_1 | H_0$) that depends on the specific test statistic that is involved and needs to take into account the correlation in test statistics because partially overlapping data are used.

The GWAS analyses were run in PLINK[1] that uses the Wald statistic to test for SNP effects. Thus, the null hypothesis that a maker has no effect is expressed as $\beta_1=0$ in the regression model $y \sim X\beta$ and tested using the Wald statistic $\hat{\beta}_1 \times \left[Var\left(\hat{\beta}_1\right)\right]^{-1/2}$. To the best of our knowledge, the distribution of the Wald statistic needed to calculate Pr($T_{(1+2)} > t$, $T_1 > c_1 | H_0$) is unknown. Therefore, below we prove that under the null hypothesis the first stage and the (combined) second stage Wald-statistics has a bivariate normal distribution with correlation:

$$\sqrt{\left[(X^TX)^{-1}\right]_{ii} / \left[(X^{(1)T}X^{(1)})^{-1}\right]_{ii}}.$$

We previously proposed[2] the *q*-value threshold of 0.1 for declaring significance. The same threshold can be used here which implies that on average we allow 10% of the SNPs that are declared to "replicate" to be false discoveries.

(1) Purcell S, Neale B, Todd-Brown K et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 2007;81:559-575.

(2) Van den Oord EJCG, Sullivan PF. False discoveries and models for gene discovery. *Trends in Genetics* 2003;19:537-542.

1 Proof

Theorem 1 *Let the predictor matrix and response vector in Stage i be $X^{(i)}$ and $y^{(i)}$, respectively, with $i = 1, 2$. Suppose we combine data from the first two stages, i.e. in the Stage 2 we apply linear regression based on the model*

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \beta + \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{bmatrix} = X\beta + \varepsilon.$$

Suppose we use Wald statistic $\widehat{\beta}_i \times \left[Var\left(\widehat{\beta}_i\right)\right]^{-1/2}$ to test the hypothesis $\beta_i = 0$. Then the correlation between the first and combined second stage Wald statistic is

$$\sqrt{\frac{\left(X^T X\right)^{-1}_{ii}}{\left(X^{(1)T} X^{(1)}\right)^{-1}_{ii}}}. \tag{1}$$

Proof. The OLR estimate $\widehat{\beta}$ of β is obtained as

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T y,$$

where $X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}$ and $y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix}$. It implies that $\left(X^T X\right)\widehat{\beta} = X^T y$. The first and second stage estimates $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ can be written in similar forms. We have that

$$\left(X^T X\right)\widehat{\beta} = X^T y = \begin{bmatrix} X^{(1)T} & X^{(2)T} \end{bmatrix} \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix} = X^{(1)T} y^{(1)} + X^{(2)T} y^{(2)} =$$

$$\left(X^{(1)T} X^{(1)}\right)\widehat{\beta}^{(1)} + \left(X^{(2)T} X^{(2)}\right)\widehat{\beta}^{(2)}.$$

From the first and the last we have that

$$\widehat{\beta} = \left(X^T X\right)^{-1}\left(X^{(1)T} X^{(1)}\right)\widehat{\beta}^{(1)} + \left(X^T X\right)^{-1}\left(X^{(2)T} X^{(2)}\right)\widehat{\beta}^{(2)} = D\widehat{\beta}^{(1)} + E\widehat{\beta}^{(2)},$$

where D and E are $p \times p$ matrices. Using the usual assumption that the errors are independent normal random variables with 0 expected value and σ^2 variance, we have that $\widehat{\beta}$ and $\widehat{\beta}^{(1)}$ have multivariate normal distribution, particularly

$$\widehat{\beta} \sim \mathbf{N}\left(\beta, \sigma^2\left(X^T X\right)^{-1}\right) \text{ and } \widehat{\beta}^{(1)} \sim \mathbf{N}\left(\beta, \sigma^2\left(X^{(1)T} X^{(1)}\right)^{-1}\right).$$

Moreover

$$Cov\left(\widehat{\beta}, \widehat{\beta}^{(1)}\right) = Cov\left(D\widehat{\beta}^{(1)} + E\widehat{\beta}^{(2)}, \widehat{\beta}^{(1)}\right) = DCov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)}\right) =$$

$$\left(X^T X\right)^{-1}\left(X^{(1)T} X^{(1)}\right) \times \sigma^2\left(X^{(1)T} X^{(1)}\right)^{-1} = \sigma^2\left(X^T X\right)^{-1} = Cov\left(\widehat{\beta}, \widehat{\beta}\right).$$

Consequently, the joint distribution of the combined Stage 2 and Stage 1 estimate of β_i is

$$\left(\widehat{\beta}_i, \widehat{\beta}_i^{(1)}\right) \sim N\left((\beta_i, \beta_i); \sigma^2 \times \begin{bmatrix} \left[\left(X^T X\right)^{-1}\right]_{ii} & \left[\left(X^T X\right)^{-1}\right]_{ii} \\ \left[\left(X^T X\right)^{-1}\right]_{ii} & \left[\left(X^{(1)T} X^{(1)}\right)^{-1}\right]_{ii} \end{bmatrix}\right).$$

Since $Var\left(\widehat{\beta}_i\right) = \sigma^2\left[\left(X^T X\right)^{-1}\right]_{ii}$ and $Var\left(\widehat{\beta}_i^{(1)}\right) = \sigma^2\left[\left(X^{(1)T} X^{(1)}\right)^{-1}\right]_{ii}$ we have that the distribution of Wald statistic is

$$\left(\frac{\widehat{\beta}_i}{\sqrt{Var\left(\widehat{\beta}_i\right)}}, \frac{\widehat{\beta}_i^{(1)}}{\sqrt{Var\left(\widehat{\beta}_i^1\right)}}\right) \sim N\left(\left(\frac{\beta_i}{\sqrt{\sigma^2\left[\left(X^T X\right)^{-1}\right]_{ii}}}, \frac{\beta_i}{\sqrt{\sigma^2\left[\left(X^{(1)T} X^{(1)}\right)^{-1}\right]_{ii}}}\right); \right.$$

$$\left. \begin{bmatrix} 1 & \sqrt{\left[\left(X^T X\right)^{-1}\right]_{ii} / \left[\left(X^{(1)T} X^{(1)}\right)^{-1}\right]_{ii}} \\ \sqrt{\left[\left(X^T X\right)^{-1}\right]_{ii} / \left[\left(X^{(1)T} X^{(1)}\right)^{-1}\right]_{ii}} & 1 \end{bmatrix}\right). \tag{2}$$

■

Remark 2 *In practice, since we do not know* $Var\left(\widehat{\beta}_i\right)$*, we approximate it by* $\widehat{Var}\left(\widehat{\beta}_i\right) = \widehat{\sigma}^2\left[\left(X^T X\right)^{-1}\right]_{ii} = \frac{\widehat{\varepsilon}^T \widehat{\varepsilon}}{n-p}\left[\left(X^T X\right)^{-1}\right]_{ii}$ *and use this approximate in Wald statistic. The statistic obtained this way,* $\widehat{\beta}_i \times \left[\widehat{Var}\left(\widehat{\beta}_i\right)\right]^{-1/2}$*, is known to have variance higher than 1, especially if the null hypothesis is not true. In particular,*

$$Var\left(\frac{\widehat{\beta}_i}{\sqrt{\widehat{Var}\left(\widehat{\beta}_i\right)}}\right) = \frac{k}{k-2} + \frac{\beta_i^2}{\sigma^2\left[\left(X^T X\right)^{-1}\right]_{ii}} \times \left[\frac{k}{k-2} - \left(\sqrt{\frac{k}{2}}\frac{\Gamma\left((k-1)/2\right)}{\Gamma\left(k/2\right)}\right)^2\right],$$

where $k = n - p$*. Also the expected value is "inflated" under the alternative hypothesis* ($\beta_i \neq 0$) *when* $Var\left(\widehat{\beta}_i\right)$ *is replaced with its estimate* $\widehat{Var}\left(\widehat{\beta}_i\right)$ *in Wald statistic:*

$$E\left(\frac{\widehat{\beta}_i}{\sqrt{\widehat{Var}\left(\widehat{\beta}_i\right)}}\right) = \frac{\beta_i}{\sqrt{\sigma^2\left[\left(X^T X\right)^{-1}\right]_{ii}}} \times \sqrt{\frac{k}{2}}\frac{\Gamma\left((k-1)/2\right)}{\Gamma\left(k/2\right)} \gtrapprox \frac{\beta_i}{\sqrt{\sigma^2\left[\left(X^T X\right)^{-1}\right]_{ii}}}.$$

Remark 3 *It is very important to note that although the variance is inflated when* $Var\left(\widehat{\beta}_i\right)$ *is replaced with its estimate* $\widehat{Var}\left(\widehat{\beta}_i\right)$ *in Wald statistic, (1) remains a very good approximation of the correlation between the first and combined second stage Wald statistic.*

Corollary 4 *Suppose we have just one predictor and we also use intercept, i.e. our matrix X has the form $X = [\mathbf{1}, \mathbf{x}]$, where $\mathbf{1}$ and $\mathbf{x}$ are vectors of the same size. Then the correlation between the first and combined second stage Wald statistic $\widehat{\beta} \times \left[Var\left(\widehat{\beta}\right)\right]^{-1/2}$ is*

$$\sqrt{\frac{\left(X^T X\right)^{-1}_{22}}{\left(X^{(1)T} X^{(1)}\right)^{-1}_{22}}} = \sqrt{\frac{n}{n\left\|\mathbf{x}\right\|^2 - \left(\mathbf{1}^T x\right)^2} : \frac{n_1}{n_1\left\|\mathbf{x}^{(1)}\right\|^2 - \left(\mathbf{1}^T x^{(1)}\right)^2}}.$$

Proof. Matrix X has the form $X = [\mathbf{1}, \mathbf{x}]$, where $\mathbf{1}$ and $\mathbf{x}$ are vectors of the same size, thus we have

$$X^T X = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{x}^T \end{bmatrix} [\mathbf{1}, \mathbf{x}] = \begin{bmatrix} n & \mathbf{1}^T x \\ \mathbf{1}^T x & \left\|\mathbf{x}\right\|^2 \end{bmatrix}$$

$$\left(X^T X\right)^{-1} = \frac{1}{n\left\|\mathbf{x}\right\|^2 - \left(\mathbf{1}^T x\right)^2} \begin{bmatrix} \left\|\mathbf{x}\right\|^2 & -\mathbf{1}^T x \\ -\mathbf{1}^T x & n \end{bmatrix}$$

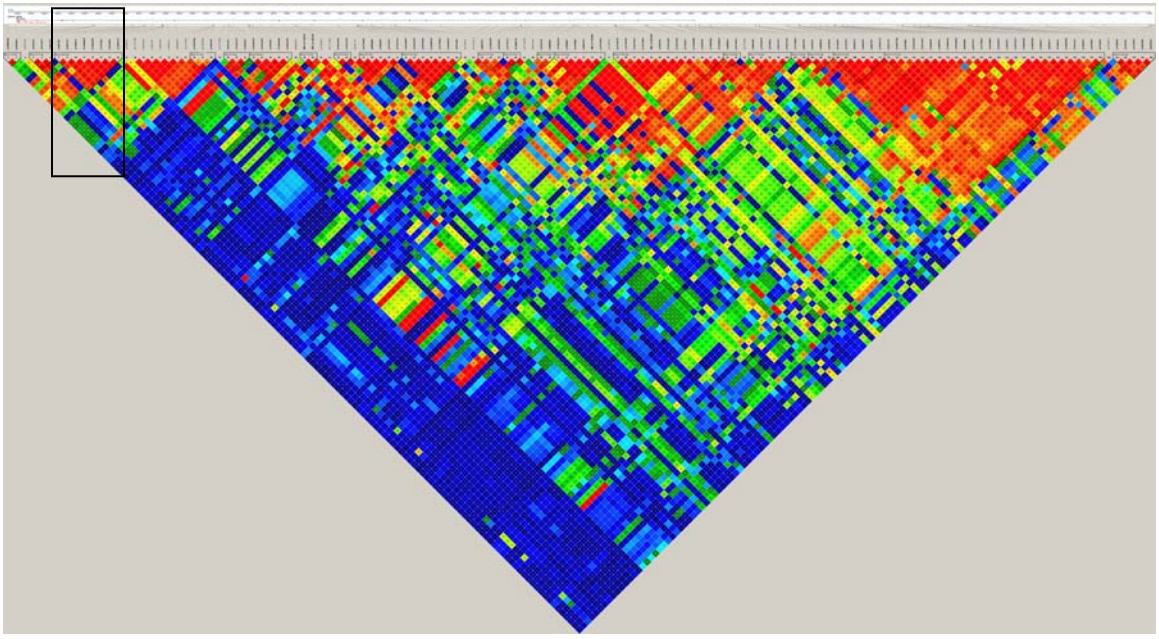
and

$$\sqrt{\frac{\left(X^T X\right)^{-1}_{22}}{\left(X^{(1)T} X^{(1)}\right)^{-1}_{22}}} = \sqrt{\frac{n}{n\left\|\mathbf{x}\right\|^2 - \left(\mathbf{1}^T x\right)^2} : \frac{n_1}{n_1\left\|\mathbf{x}^{(1)}\right\|^2 - \left(\mathbf{1}^T x^{(1)}\right)^2}}.$$

■

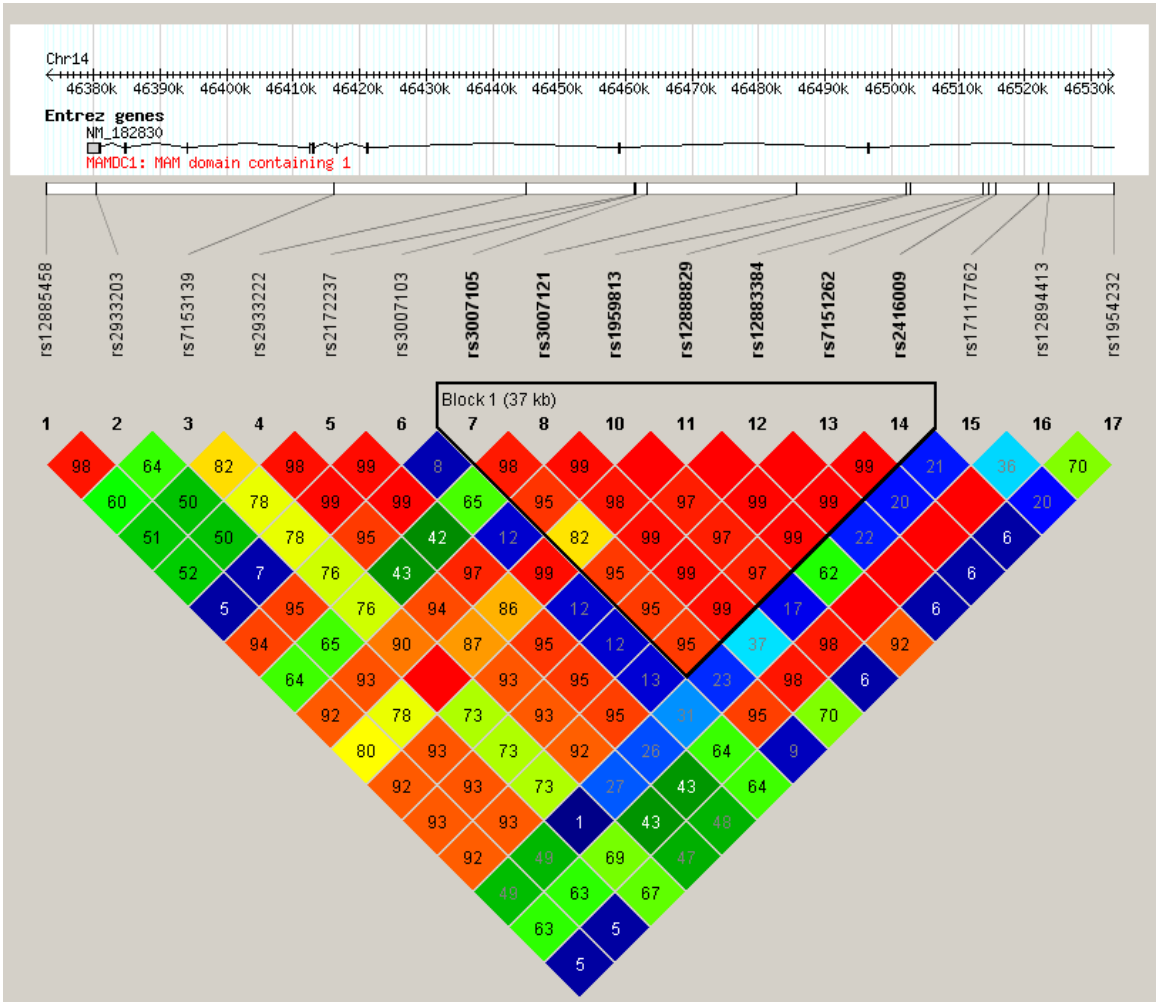
4. LINKAGE DISEQUILIBRIUM (D') PLOT FOR MDGA2 GENE

Supplemental Figure 1 shows the Linkage Disequilibrium (D') across entire MDGA2 gene. Box is region containing 4 associated SNPs.



Supplemental Figure 1: Linkage Disequilibrium (D') across entire MDGA2 gene.

Supplemental Figure 2 below shows Linkage Disequilibrium (D') in region of association within MDGA2 gene.



Supplemental Figure 2. Linkage Disequilibrium (D') in region of association within MDGA2 gene.

Associated SNP are markers 7, 10, 12 and 14. Block was defined manually to highlight region of high marker to marker LD and covers exon 10. There are two isoforms of MDGA2 which have alternate non overlapping starting exons. If all possible exons are considered, the associated region encompasses exon 11.