A transcriptome wide association study in conjunction with fine-mapping identifies four potentially causal genes for neuroticism

Benjamin Russoniello, Hunter Melton

December 12 2020

Abstract

Previous genome-wide association studies have identified 1,716 disparate variants that are associated with neuroticism at a genome-wide significant level; however, the genes accountable for this association have only briefly been examined. In an effort to identify probable causal genes, we conducted a transcriptome-wide association study on neuroticism in 270,059 subjects of European ancestry. We predicted gene expression in brain cortex tissue by using data from the Genotype-Tissue Expression project. Out of 1,068 genes evaluated in the model, 45 were identified as significant using a false discovery rate corrected p-value of 2.042×10^{-3}. We further performed fine-mapping to limit the impact of linkage disequilibrium on our results, limiting us to a potentially causal set of 4 genes across 22 chromosomes.

Introduction

Neuroticism, identified as one of the five factors of personality in John Digman's seminal 1990 paper, is of increasing concern in public health disciplines [1]. Defined as "relatively stable tendencies to respond with negative emotions to threat, frustration, or loss", neuroticism is typically marked by intense negative emotional reactions to what others may find only mildly concerning [4]. Given that it is highly correlated with substance abuse, depression, anxiety, and a number of other maladies, it is of no surprise neuroticism is drawing increasing awareness in public health circles. Past genome-wide association studies (GWAS) on neuroticism have identified some 1,716 significant variants, but little work has been done on unifying this with gene expression data, and none attempting to determine causal genes.

Even with large sample size, in GWAS, it can be difficult to identify individual SNPs as significant given the minimal effect size of many variants. As a result, gene-based methods, which bundle many individual variants together into a single unit, can be employed to increase power [8]. Transcriptome-wide association studies (TWAS), which explores the association between disease and predicted gene expression, are exactly such a method. We perform a TWAS on neruoticism employing the TWAS/FUSION method [2] to analyze the GWAS summary statistics of 270,059 subjects of European descent [3].

GWAS or TWAS analysis may be extended by fine-mapping, which seeks to constrain a minimum set of causal variants to explain association. Linkage Disequilibrium (LD) can induce signal at non-causal genes, resulting in incorrectly calling these genes causal. Fine-mapping guards against this, and we use the FOCUS [5] method to perform fine-mapping on the significant genes identified by the neuroticism TWAS.

Methods

TWAS/FUSION

This analysis is composed of two steps. First, a model for predicting gene expression levels in brain cortex tissue was created through the use of transcriptome data from the Genotype-Tissue Expression (GTEx) Project. Then, while accounting for LD using reference data from the 1000 Genomes Project, we employed TWAS/FUSION [2] to analyze the association between predicted gene expression and neuroticism risk for the aforementioned 270,059 individuals with European ancestry in the GWAS [3]. A more in-depth discussion of the method can be examined in the original paper from Gusev et. al, but we will briefly describe the second step. Let $\hat{G}$ be the imputed gene expression for the GWAS summary data, so $\hat{G} = X\hat{\beta}$, where X is the GWAS summary data and $\hat{\beta}$ comes from the gene expression model in the first step. Then, letting Y be the trait of interest, neuroticism, we have:

$$Y = \gamma\hat{G} + \epsilon = \gamma X\hat{\beta} + \epsilon = \gamma_1 X_1\hat{\beta}_1 + ... + \gamma_p X_p\hat{\beta}_p$$

We want to test $\gamma = 0$, which is the same as testing $(\gamma_1, ..., \gamma_p) = 0$. This process results in the outcomes provided below.

Fine-mapping with FOCUS

After receiving the results from TWAS, FOCUS [5] was used to fine-map the results, in order to find genes that were responsible for the association signal across a chromosome. This is done via a Bayesian approach. If c_i represents gene i being causal, then,

$$c_i \sim Binomial(10^{-3})$$

Further, we assume the model,

$$y = X\beta + G\alpha + \epsilon$$

, where y is the trait expression, X is our standardized SNP expression, β is the pleiotropic effects of SNPs on the expression y, G is the matrix of gene expression data , and α is causal effects for the genes. G is estimated with $\hat{G}$, where $\hat{G} = XW$.

Mansuco et al. show that under this model, a probability distribution of the TWAS test statistics conditional on gene expression data and LD weights can be computed. Finally, using Bayes rule, a formula is given to compute the posterior distribution of any set of causal genes $\mathbf{c}$. Then, we are able to find 100∗p% credible sets of genes, where $0 < p < 1$. p is typically chosen to be high, and by default in the FOCUS software it is set to be .9. A $100 * p\%$ credible set contains a causal gene with probability p. Additionally, a posterior inclusion probability (PIP) is calculated for each gene. A PIP gives the probability that the gene is included in the true model given above. In this sense, fine-mapping can be thought of as a variable selection problem, and FOCUS is a technique to conduct that variable selection.

Results

From the TWAS analysis, 45 of 1,068 genes were identified as significant using a false discovery rate-corrected p-value of $2.042 * 10^{-3}$, which corresponds to a Z-score of 3.084. These 45 genes are reported in Table 1. 30 of these genes had a higher expression associated with a higher risk of neuroticism, and the remaining 15 had a higher expression associated with a lower risk of neuroticism. If we use a Bonferroni corrected p-value of $4.68 * 10^{-5}$, only 19 of these genes maintain their significance. All 45 significant genes had been previously identified in TWAS studies, so we are quite confident in these results.

We validated this analysis through repetition. Using a smaller GWAS [7] of 168,105 subjects of European descent, we fully replicated our analysis. Through this, 19 of the previous 45 genes were identified to be transcriptome-wide significant at a false discovery rate-corrected p-value.

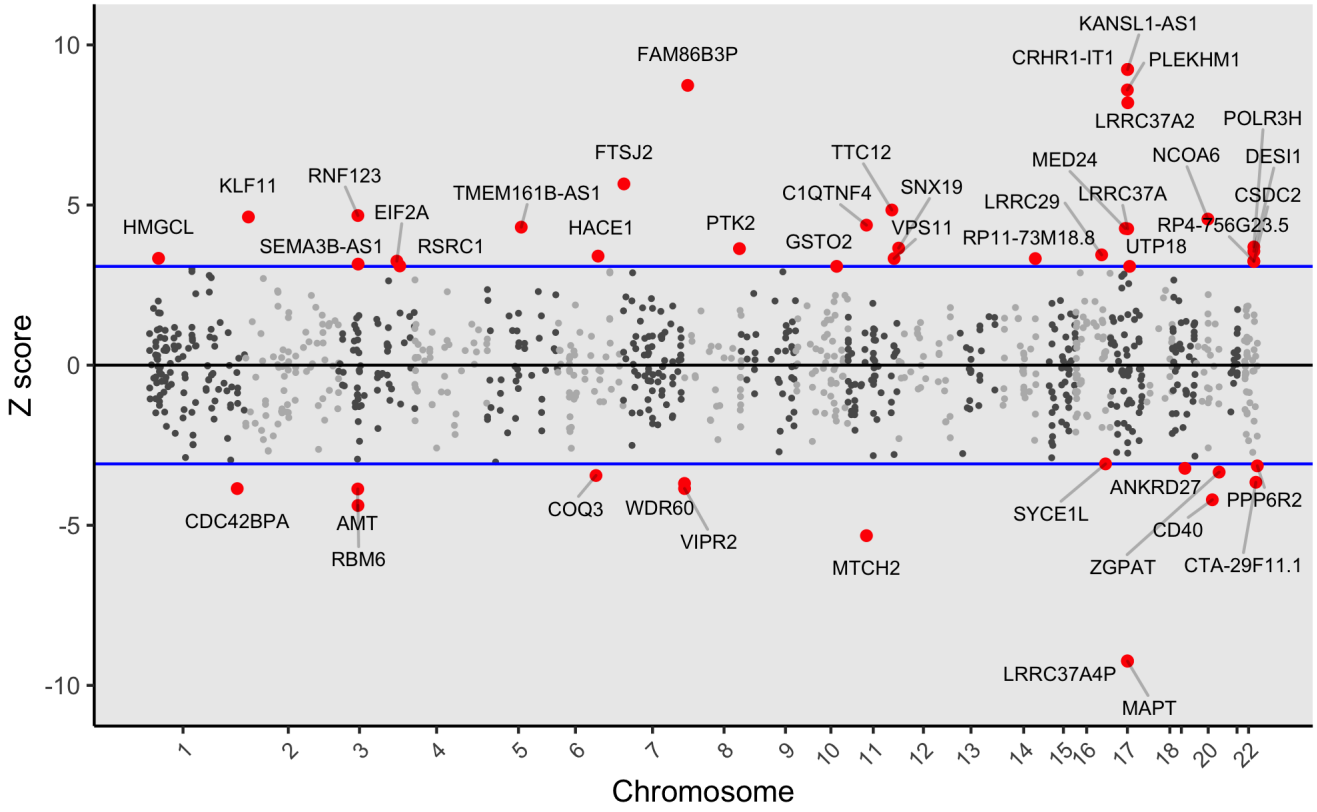

Figure 1: Manhattan plot of association results from TWAS.

The blue lines correspond to a Z score of ±3.084.

The results of fine-mapping with FOCUS are summarized in Table 2. Each gene in the table was contained in the 90% credible set pertaining to its chromosome location. One notable result from this is the identification of the MAPT gene. The MAPT gene has been shown to have significant association with the development of Alzheimers [6]. While there is no direct association between neuroticism and Alzheimers, both are known to affect mood and decision-making skills through pathways in the brain.

Table 1: Significant genes from TWAS

Gene	**Chr**	**Z Score**	**P value**	**GWAS SNP**[1]	**Heritability**[2]
HMGCL	1	3.33604	0.00085	rs10799802	0.215
CDC42BPA	1	-3.85475	0.000116	rs6667260	0.254
GSTO2	10	3.085	0.00204	rs7909129	0.3699
C1QTNF4	11	4.372	1.23e-05	rs11039149	0.256
MTCH2	11	-5.325	1.01e-07	rs11039149	0.211
TTC12	11	4.8449	1.27e-06	rs4534613	0.481
VPS11	11	3.33	0.000868	rs17075	0.399
SNX19	11	3.6548	0.000257	rs2131535	0.363
RP11-73M18.8	14	3.328	0.000875	rs8015712	0.353
LRRC29	16	3.444	0.000573	rs12927959	0.532
SYCE1L	16	-3.084	0.002042	rs682517	0.544
MED24	17	4.269	1.96e-05	rs12601929	0.361
PLEKHM1	17	8.594	8.38e-18	rs17689882	0.207
LRRC37A4P	17	-9.234	2.61e-20	rs17689882	0.513
CRHR1-IT1	17	9.233	2.63e-20	rs17689882	0.285
MAPT	17	-9.246	2.33e-20	rs17689882	0.137
KANSL1-AS1	17	9.234	2.61e-20	rs17689882	0.391
LRRC37A	17	4.258	2.06e-05	rs17689882	0.596
LRRC37A2	17	8.194	2.52e-16	rs199439	0.459
UTP18	17	3.084	0.00204	rs1263956	0.274
ANKRD27	19	-3.2231	0.00127	rs4805758	0.316
KLF11	2	4.627	3.71e-06	rs4669520	0.422
NCOA6	20	4.562	5.07e-06	rs7265992	0.1426
CD40	20	-4.2043	2.62e-05	rs2425752	0.3593
ZGPAT	20	-3.34	0.000838	rs2273487	0.4074
RP4-756G23.5	22	3.253	0.001142	rs2024566	0.099
POLR3H	22	3.691	0.000223	rs2024566	0.267
CSDC2	22	3.235	0.001216	rs2024566	0.309
DESI1	22	3.5566	0.000376	rs2024566	0.17
CTA-29F11.1	22	-3.6578	0.000254	rs139590	0.324
PPP6R2	22	-3.1453	0.001659	rs9616393	0.215
AMT	3	-3.8675	0.00011	rs34759087	0.106
RNF123	3	4.672	2.98e-06	rs4625	0.232
RBM6	3	-4.3822	1.18e-05	rs4625	0.214
SEMA3B-AS1	3	3.1553	0.0016	rs695238	0.67
EIF2A	3	3.248	0.00116	rs6799014	0.455
RSRC1	3	3.098	0.00195	rs3845980	0.429
TMEM161B-AS1	5	4.309	1.64e-05	rs16903275	0.3542
COQ3	6	-3.447	0.000567	rs4431442	0.202
HACE1	6	3.405	0.000662	rs457286	0.293
FTSJ2	7	5.659	1.52e-08	rs11764590	0.168
WDR60	7	-3.693	0.000221	rs6979985	0.32
VIPR2	7	-3.854	0.000116	rs6979985	0.349
FAM86B3P	8	8.7362	2.41e-18	rs777709	0.136
PTK2	8	3.638	0.000275	rs11996715	0.223

[1]Most significant GWAS SNP in locus. [2]Heritability of gene.

Table 2: FOCUS Results: Genes contained in 90% credible sets

Gene	Chr	PIP
C1QTNF4	11	.883
TTC12	11	.882
MAPT	17	.85
POLR3H	22	.116

Discussion and Conclusions

Using FUSION, we were able to conduct a TWAS on recently published GWAS data on neuroticism. This enabled us to link significant variants to genes that are associated with neuroticism. Finding genes that have an association can facilitate the production of therapies, as genes can be mapped to protein production that affect brain function. Additionally, using FOCUS, we were able to narrow our focus down to 4 genes that are potentially causal in expression of neuroticism. If genetic models of neuroticism are considered in the future, the 4 genes identified by FOCUS can be thought of as a prioritized set of variables worthy of consideration in such a model.

Naturally, we have identified a few candidate areas to expand this work in the future. Primarily, we plan to expand to more tissues than solely the brain cortex for the gene expression. This will likely help to identify novel significant genes and loci associated with neuroticism risk, as well as illuminating more causal genes through FOCUS. Furthermore, additional validation is necessary for all of the given results, preferably with a large amalgamation of independent GWAS data. Finally, as stated above, an investigation of the proteins associated with the identified causal genes may be useful in the eventual development of therapies for neuroticism.

References

[1] John M. Digman. "Personality structure: emergence of the five-factor model". In: *Annual Review of Psychology* (1990). ISSN: 00664308. DOI: 10.1146/annurev.ps.41.020190.002221.

[2] Alexander Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature Genetics* (2016). ISSN: 15461718. DOI: 10.1038/ng.3506.

[3] David W Hill. "Genetic contributions to two special factors of neuroticism are associated with affluence, higher intelligence, better health, and longer life". In: *Molecular Psychiatry* 25.11 (2020). DOI: 10.1038/s41380-019-0387-3.

[4] Benjamin B. Lahey. "Public Health Significance of Neuroticism". In: *American Psychologist* (2009). ISSN: 0003066X. DOI: 10.1037/a0015309.

[5] Nicholas Mancuso et al. "Probabilistic fine-mapping of transcriptome-wide association studies". In: *Nature Genetics* (2019). ISSN: 15461718. DOI: 10.1038/s41588-019-0367-1.

[6] A.J. Myers et al. "The H1c haplotype at the MAPT locus is associated with Alzheimer's disease". In: *Human Molecular Genetics* 14.16 (July 2005), pp. 2399–2404. ISSN: 0964-6906. DOI: 10.1093/hmg/ddi241. eprint: https://academic.oup.com/hmg/article-pdf/14/16/2399/1968226/ddi241.pdf. URL: https://doi.org/10.1093/hmg/ddi241.

[7] Patrick Turley et al. "Multi-trait analysis of genome-wide association summary statistics using MTAG". In: *Nature Genetics* (2018). ISSN: 15461718. DOI: 10.1038/s41588-017-0009-4.

[8] Lang Wu et al. "A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer". In: *Nature Genetics* (2018). ISSN: 15461718. DOI: 10.1038/s41588-018-0132-x.