Biomedical Data Science

Department of Biostatistics and Medical Informatics

UNIVERSITY OF WISCONSIN

SCHOOL OF MEDICINE AND PUBLIC HEALTH


---

# STAT/BMI-620 Project Report

## GWAS Analysis on Income

**Done by: Yushin Wei, Arnav Deshpande & Jiren Lu**

# Table of Contents

# 1. About the data:

We analyzed the sumstats from a published research: *Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income.* The sample comes from participating individuals from the UK Biobank. They initially had around 332050 people in it. The phenotypic trait being studied is self-reported income on a pretty crude ordinal scale: Answer 1 if income less than £18,000, 2 if between £18,000–£29,999, 3 if between £30,000–£51,999, 4 if between £52,000–£100,000 and 5 if more than £100,000. Individuals answering “do not know” or “prefer not to say” were excluded from the dataset. In the end, the sample includes a total of 286301 people with 138425 male and the rest female. The age range is 39-73 with a mean of 56.5 years.

# 2. Methods used

## a. Data Visualization

### i. Manhattan plot & QQ Plot:

A Manhattan plot was constructed using R, with the code we referenced in homework 3.A quantile-quantile (qq) plot tries to match the observed p-value distribution from GWAS to the expected distribution under the null, this was made using R.

### ii. LocusZoom Plot:

We used the locuszoomr package available on CRAN. It is similar if not better compared to those obtained from U-Mich's locuszoom.org. LocusZoom is a regional association plot designed to narrow down to a specific locus to show the association signals of the SNP in that region, the LD between SNPs and genes located near that region.

## b. Linkage disequilibrium score (LDSC)

### i. Heritablity and Enrichment plot

We used ldsc software to analyze the heritability and enrichment data of income using the reference pipeline from homework 4 of this class. The enrichment plot was then generated using R.

### ii. Pairwise genetic correlation

We chose male pattern baldness, ADHD, leisure computer usage, lefthandedness, and cannabis usage to analyze the pairwise genetic correlation. The pipeline also comes from homework 4 and software ldsc was used. A p-value of 0.001 was the cut off for significance.

## c. Performing TWAS - Fusion

We used Transcriptome Wide Association Study (TWAS) to examine the RNA products of tissues and gives us a window to look at the genes expressed and their expression level. We used FUSION (Gusev et al. “Integrative approaches for large-scale transcriptome-wide association studies” 2016 Nature Genetics) stands for Functional Summary-based Imputation. It is a framework/toolset that can be used to effectively perform TWAS. This software is adaptive enough to take information from large datasets using imputation. Imputation methods use some sort of a Bayesian sparse linear-mixed model.
Find the TWAS code here.

## d. Mendelian Randomization - IVW

We use Mendelian Randomization to test whether income causally affects five traits. We filter genome-wide significant SNPs ($p < 5\times10^{-8}$) from the income GWAS as exposures, match them with each outcome GWAS, and run IVW MR to estimate causal effects.

# 3. Results

(Find the data visualization code here)

## a. Results of Data Visualization:

### i. Manhattan plot & QQ Plot:

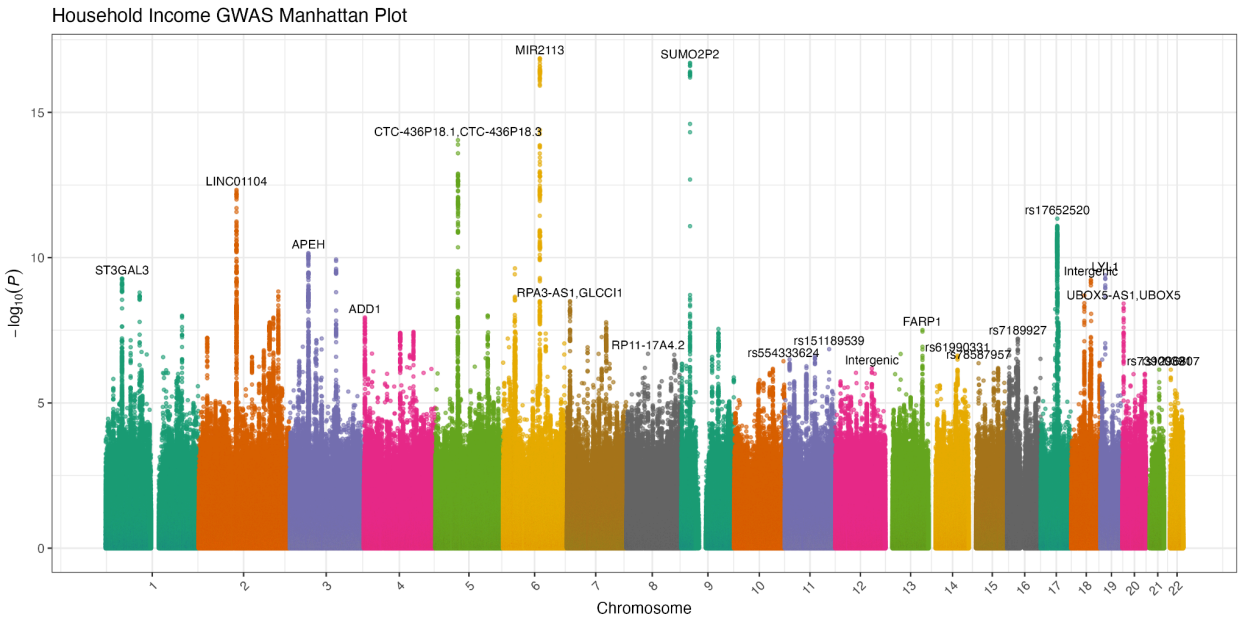



Figure 1, the Manhattan plot of income.

Figure 1 is the Manhattan plot for the income phenotype. The graph is quite polygenic in nature. Many loci are non coding/intergenic. Significant loci seen across chromosomes 1, 2, 3, 5, 6, 9, 17 and 18. Several genes observed such as ST3GAL3 on CHR 1, associated with educational attainment, intelligence and cognition. LINC01104 on CHR 2 is also associated with intelligence and cognitive ability. Two most significant loci on CHR 6 and CHR 9 with p-values exceeding the order of $10^{-17}$. SUMO2p2 is a pseudogene on CHR 9 which serves as a biomarker for LINC01239. There are a few studies showing this to be associated with educational attainment and cognitive ability. Our top hit is microRNA 2113, or MIR2113, which I shall elaborate on with detail in the LocusZoom plot section.

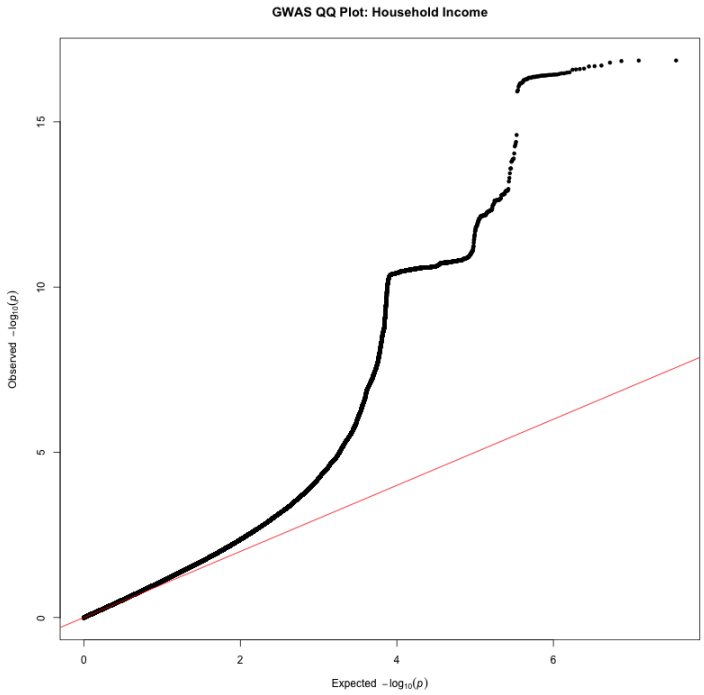



Figure 2, QQ plot of the GWAS

Figure 2 shows the quantile-quantile (qq) plot for the  income phenotype. It's a decent fit to p-values less than $10^{-3}$ but then deviates super haphazardly. Largest deviations of ~$10^{-17}$ do match the manhattan plot. Step like behavior indicates quite a bit of genomic inflation, which is corroborated by our LDSC. This could mean that: 1, **Household income is highly polygenic with many SNPs having small effects. 2,** Confounding or stratification (not the case since our LDSC intercept is close to 1)

ii. LocusZoom Plot:

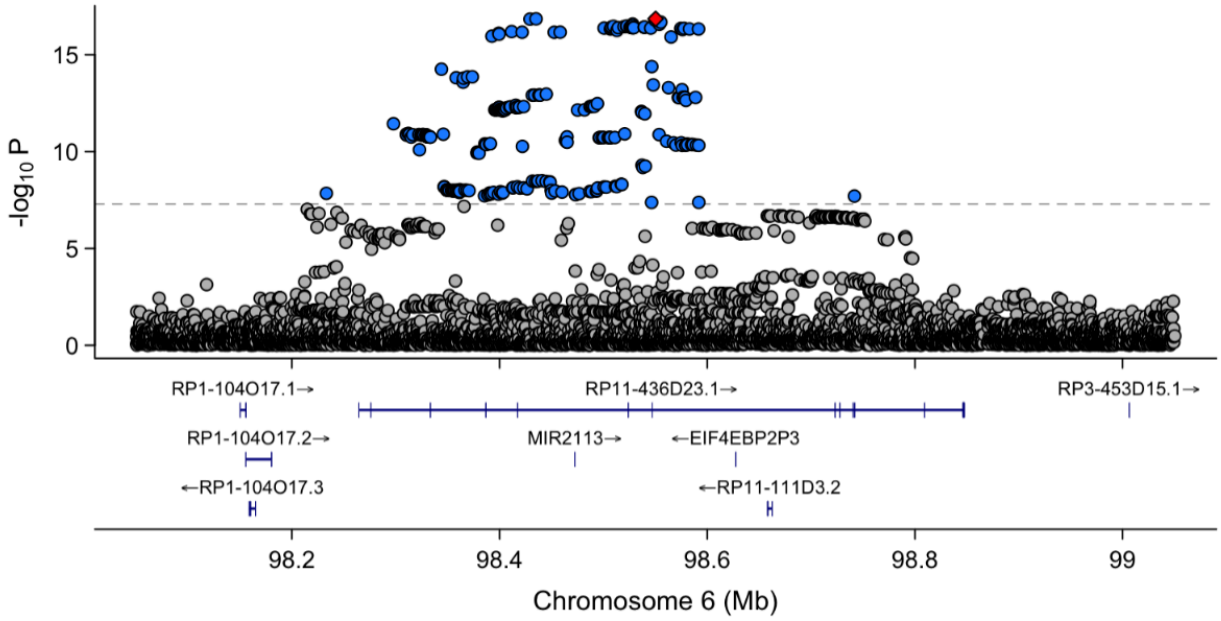



Figure 3, LocusZoom plot tells us a story about the most significant genes across the data.

The most significant and important gene out of all the ones above is MIR2113, or better known as microRNA2113. MIR2113 is heavily associated with cognitive ability and intelligence. The cytogenetic region it is located at is called 6q16. A microdeletion of 6q16, where a segment of a chromosome gets deleted and the two ends which originally were attached to the segment instead get fused to one another, causes a rare Prader-Willi like syndrome which includes a series of severe organ and neurological defects including global developmental delay in the body, a shrunken brain, anomalies in hands, feet, heart, kidney and other organs. This further corroborates the fact that this region as a whole is vital for maintaining your cognition. The other relevant gene, EIF4EBP2P3, is also highly linked with intelligence and educational attainment. The gene stands for: **E**ukaryotic Translation **I**nitiation **F**actor **4E** **B**inding **P**rotein **2** Pseudogene **3.**

## b. Results of LDSC

Find results here
Linkage disequilibrium score (LDSC) regression showed that the mean $\chi 2$ statistic was 1.4498, Lambda GS was 1.3615, and the intercept of the LDSC regression was 1.0358 (SE = 0.0081). These statistics indicate that there is a high polygenity signal rather than residual stratification or confounding. The LDSC regression estimate of the heritability of household income was 0.0745 (SE = 0.0032), the $\chi 2$ of heritability is 494.9, corresponding to a very small p-value. This indicates that income is heritable, but the heritability is very low.
The same method was used to process the data for 5 other traits, listed in table 1

| | Correlation | p-value |
|---|---|---|
| ADHD | -0.5115 (0.0265) | 7.56E-83*** |
| Baldness | -0.0013 (0.0239) | 0.9561 |
| Cannabis usage | -0.4237 (0.0407) | 2.00E-25*** |
| computer usage | 0.469 (0.023) | 4.58E-92*** |
| lefthandedness | 0.0431 (0.0434) | 0.32 |

Table 1, the genetic correlation of income with 5 other traits

We found higher income is negatively associated with ADHD and cannabis usage, and is positively associated with leisure computer usage. There is no statistical significant association for income with baldness and lefthandedness.

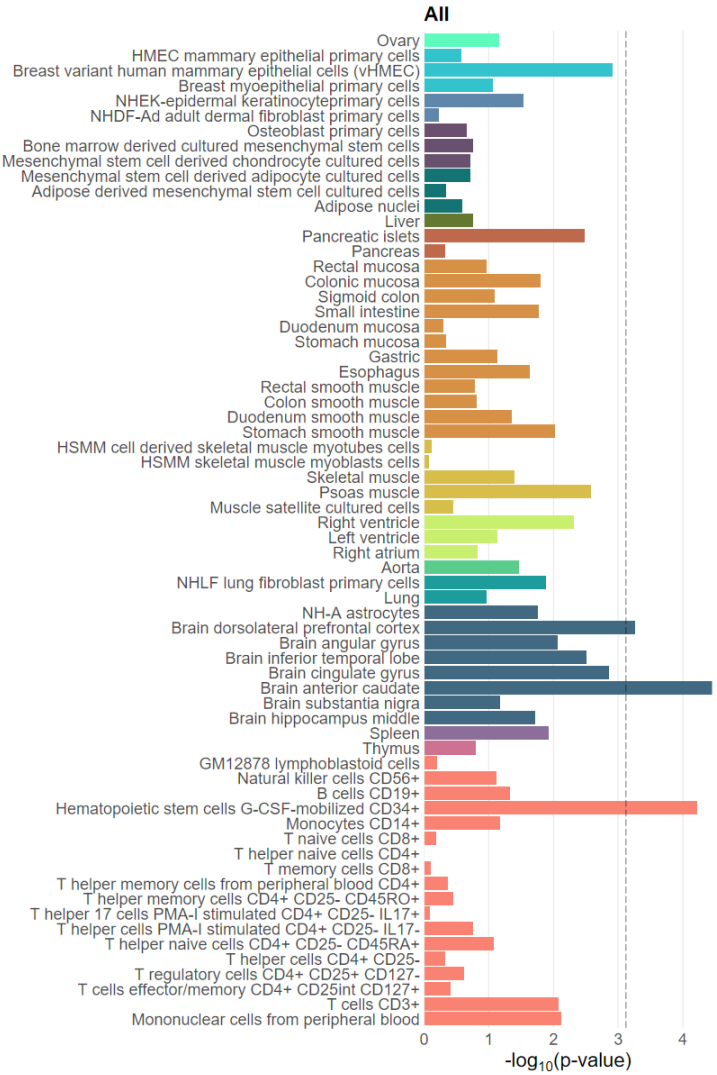
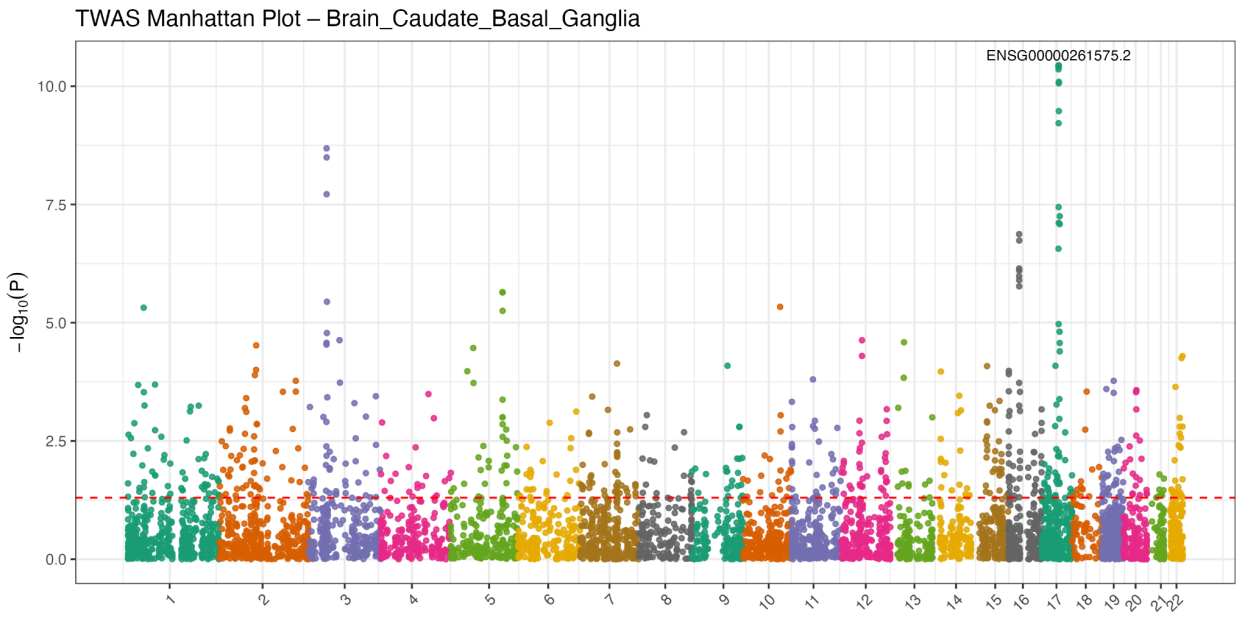



Figure 4, the enrichment of the genes associated with income in different tissue types.

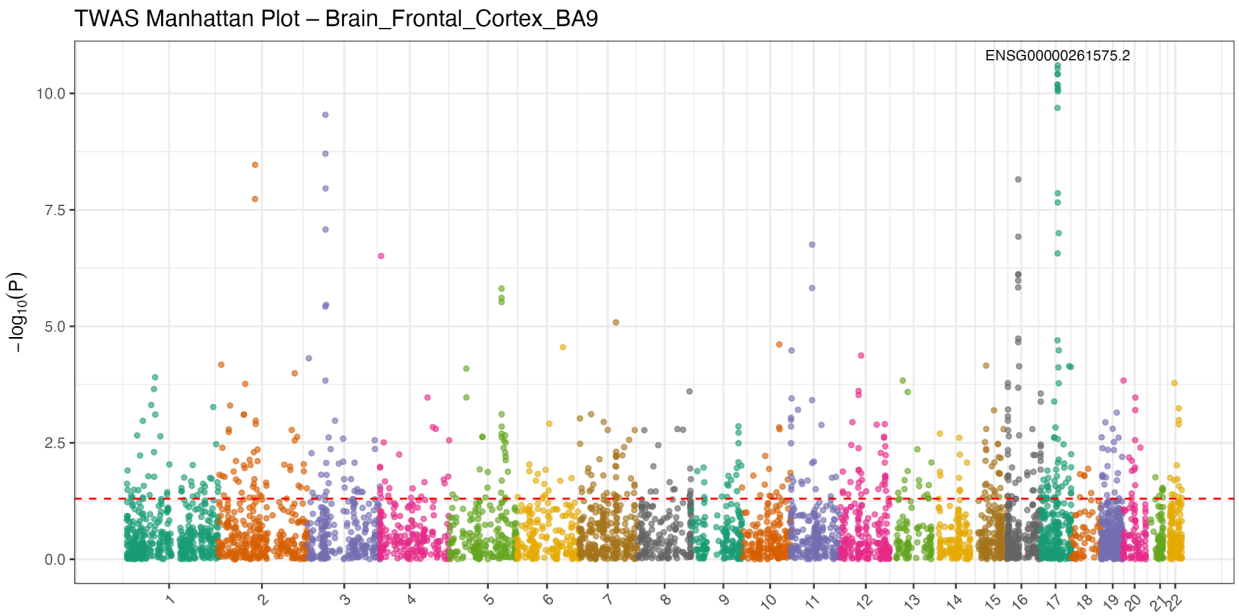
Figure 4 is the enrichment plot. We found income associated genes have the high enrichment in brain related tissues (marked as dark blue). The highest enrichment is in frontal caudate, which plays a critical role in motor control, planning, learning, and reward processing.

## c. Results of TWAS - FUSION

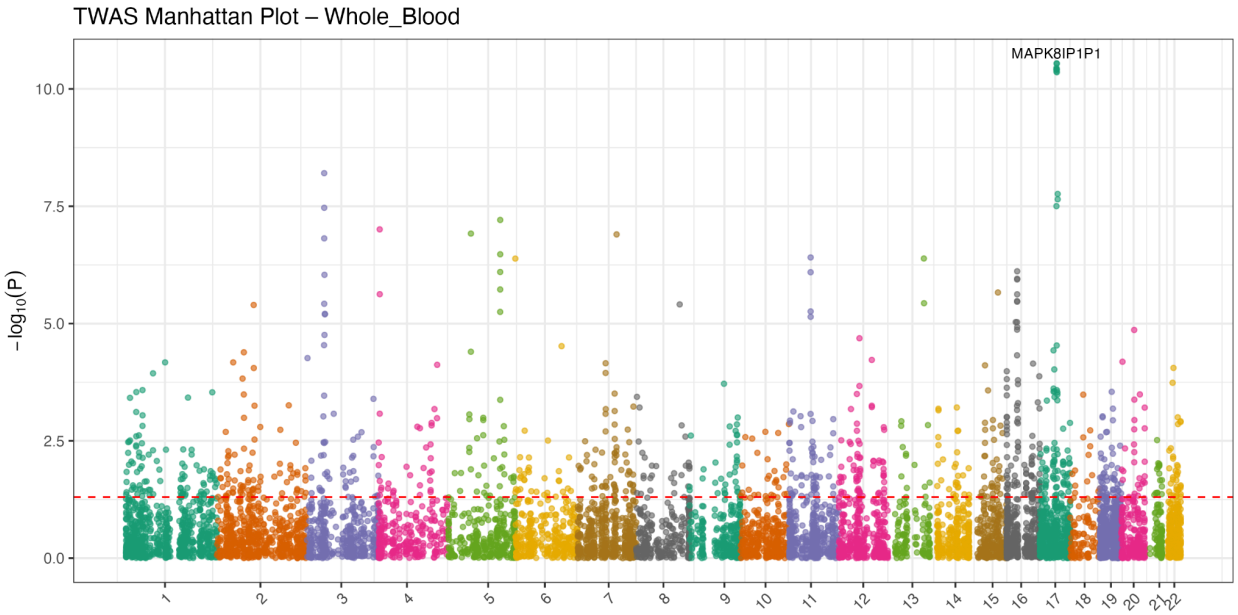
Since our enrichment plot indicated that the brain region and STEMcell regions were significant, we went ahead and ran TWAS on most of the brain tissues that were available on the website, along with whole blood since no weights for STEM cells were available. Here are the manhattan plots from the same:



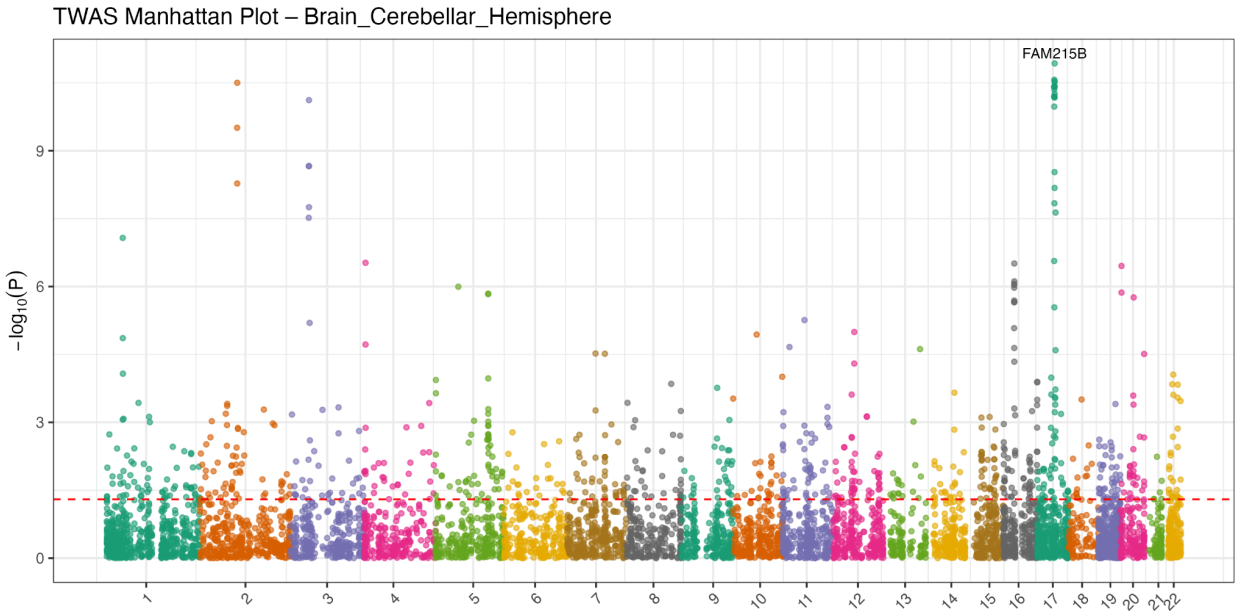

- ENSG00000261575.2 is the most relevant gene. It is a pseudogene named C17orf58. It serves as a biomarker for the second gene, ARHGAP27, which is actually significant.
- ARHGAP27 is a Rho-GTPase 27, which is responsible for your neuronal development and synaptic functions.
- It's very closely associated with cortical thickness, educational assessment and neuroticism. The more one scrutinizes their own mistakes, the more stress you take, the more money you make.

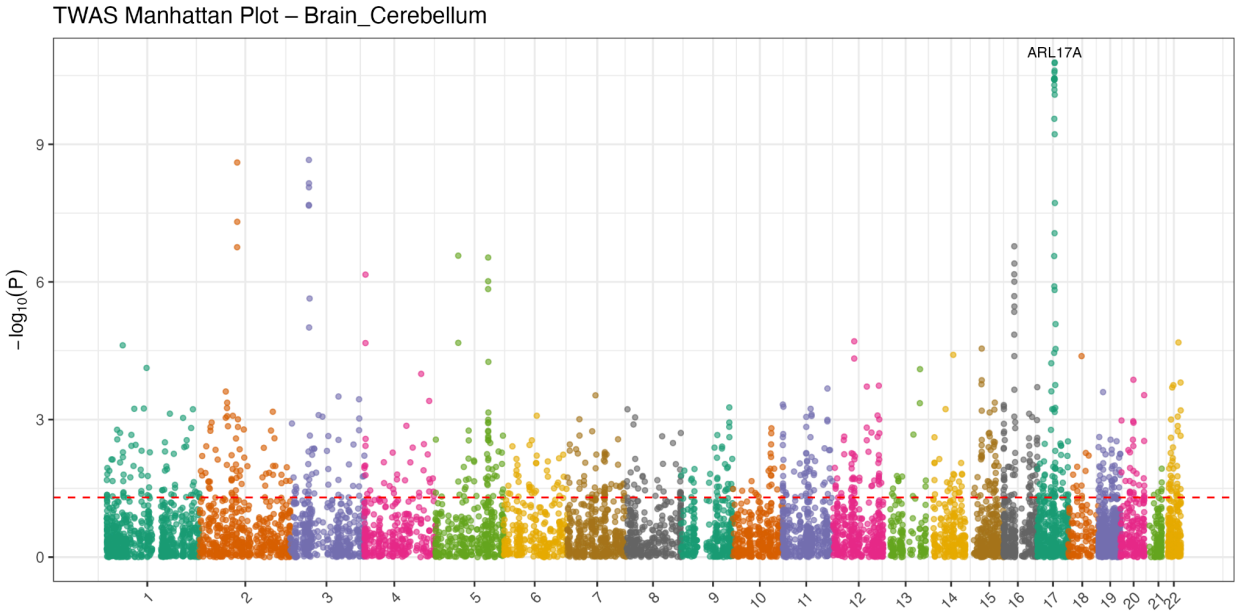



- I analyzed Frontal Cortex Brodmann Area 9.
- BA9 has the function of memory coding and inferential reasoning.
- Gave me the same top pseudogene-actual gene combo of C17orf58 and ARHGAP27.

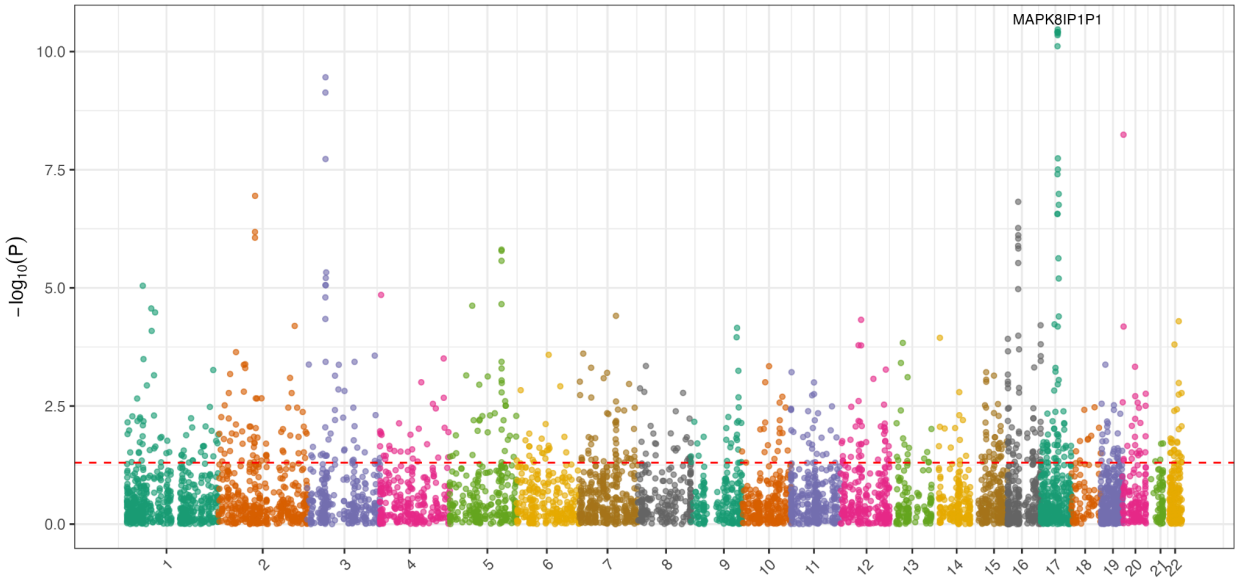


- The Whole Blood analysis yielded the top pseudogene MAPK8IP1P1 and gene combination LINC02210, which is associated with intelligence, neuroticism, highest math class taken and self-reported EA.
- Some other analyses of brain regions:

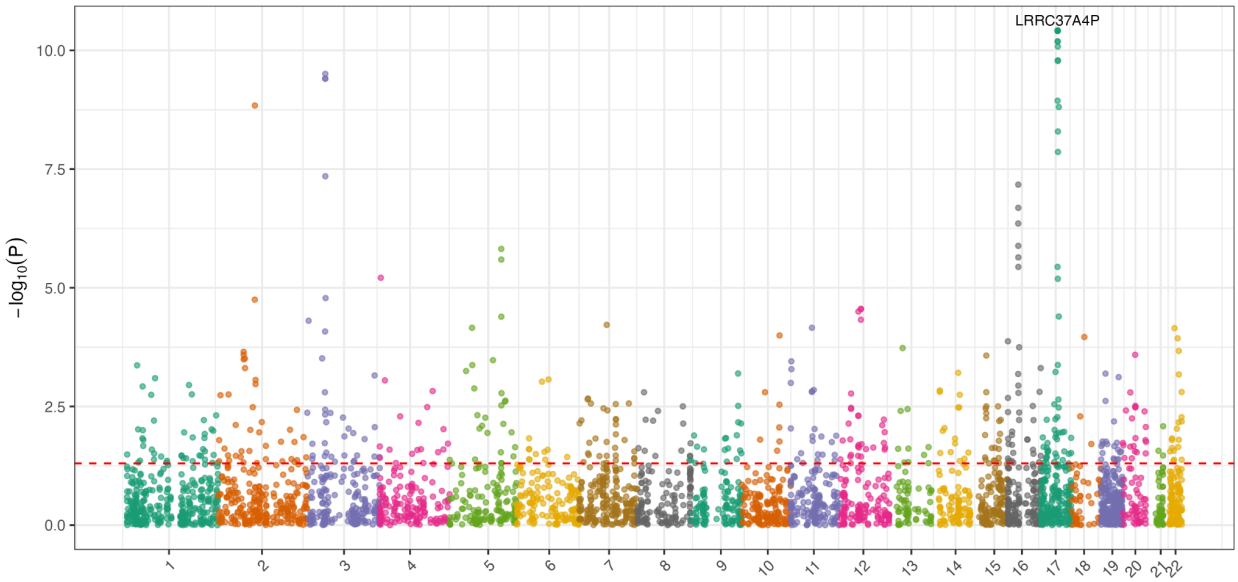


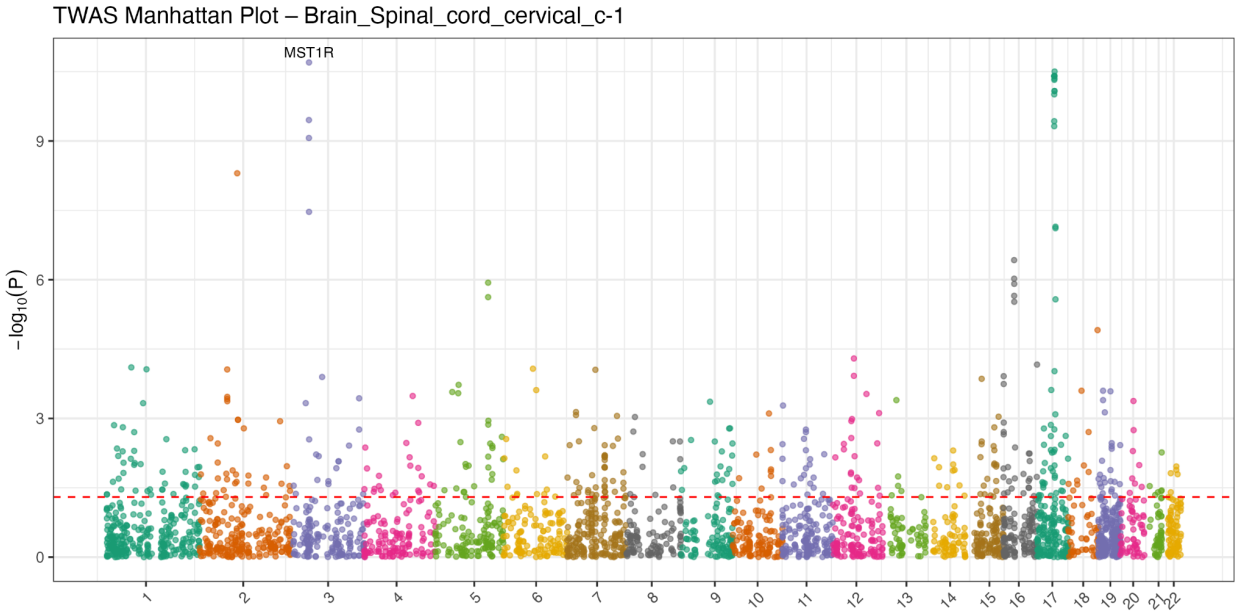



TWAS Manhattan Plot – Brain_Cortex




TWAS Manhattan Plot – Brain_Hypothalamus






| Tissue | Top Gene | Chromosome | Significance (-log10(P)) |
|---|---|---|---|
| BrainCerebellarHemisphere | **FAM215B** | chr17 | ~9.5 |
| Brain Cerebellum | **ARL17A** | chr17 | ~9.5 |
| Brain Cortex | **MAPK8IP1P1** | chr17 | ~10.5 |
| Brain Hypothalamus | **LRRC37A4P** | chr17 | ~10.5 |
| BrainSpinalCordCervicalC-1 | **MST1R** | chr3 | ~9.5 |

Table 2, TWAS significant area. FAM215B is also known as ARL17A-IT1. ARL17A is related with brain volume, body height and brain connectivity attribute. MAPK8IP1P1 is the same pseudogene found in the whole blood TWAS, and associated with the same actual gene LINC02210. LRRC37A2 is associated with brain connectivity attribute, reaction time

measurement and intelligence. MST1R is associated with self-reported EA, sedentary behavior, leisure screentime and cognitive function measurement.

## d. Results of MR - IVW

The IVW MR results are as follows.

| Trait | num SNPs | Beta | SE | p-value |
|---|---|---|---|---|
| ADHD | 2977 | -1.509 | 0.0166 | $< 1 \times 10^{-300}$ |
| Baldness | 3149 | 4.627 | 0.0683 | $< 1 \times 10^{-300}$ |
| Cannabis Dependence | 2973 | -0.066 | 0.0026 | $2.96 \times 10^{-145}$ |
| Left Handedness | 3202 | 0.620 | 0.0081 | $< 1 \times 10^{-300}$ |
| Leisure Computer Use | 3203 | 0.197 | 0.0085 | $3.13 \times 10^{-119}$ |

Table 3, the output of Mendelian randomization-IVW

The results in Table 3 indicate that income has statistically significant causal effects on all five traits analyzed. Higher income is associated with a substantial reduction in ADHD risk and cannabis dependence, and is positively associated with baldness, left-handedness, and leisure computer use. All associations are highly significant with p-values far below conventional thresholds.

Although LDSC found no genetic correlation between income and baldness or left-handedness, MR showed strong causal effects. This suggests income may influence these traits through non-genetic pathways. For baldness, this could involve lifestyle or health factors linked to income. For left-handedness, higher income may relate to reduced social pressure or differences in early-life environments that affect handedness.

# 4. Conclusion

Our analyses spanned multiple genomic and statistical approaches, including data visualization, LDSC regression, TWAS using FUSION, and Mendelian Randomization. The Manhattan and QQ plots revealed a highly polygenic architecture for income, with significant loci across multiple chromosomes. Notably, genes such as MIR2113 and ARHGAP27 were identified as key contributors, aligning with prior findings linking these regions to cognitive and educational traits.

LDSC results supported a modest but significant heritability of income, with enrichment predominantly in brain tissues—particularly the frontal caudate—underscoring the neurological underpinnings of socioeconomic traits. Genetic correlations suggested a complex interplay between income and neuropsychiatric, behavioral, and lifestyle traits. Mendelian Randomization further highlighted potential causal pathways, revealing that higher income may reduce risks for ADHD and cannabis dependence while increasing the likelihood of baldness, left-handedness, and leisure computer use.

Overall, our findings reinforce the polygenic and neurologically rooted nature of income, while also suggesting broader behavioral and lifestyle associations. These results contribute to a growing understanding of the genetic influences on socioeconomic outcomes and point to potential avenues for future interdisciplinary research.

# 5. Member contributions

**Yushin Wei: LDSC, enrichment and Pairwise Correlation**
**Arnav Deshpande: TWAS, LocusZoom, Manhattan & QQ Plot**
**Jiren Lu: Mendelian Randomization - IVW**

**Equal contributions for presentation and report by all members.**

# Reference


**Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." *Nature genetics* 47, no. 3 (2015): 291-295.**

**Burgess, Stephen, Adam Butterworth, and Simon G. Thompson. "Mendelian randomization analysis with multiple genetic variants using summarized data." *Genetic epidemiology* 37, no. 7 (2013): 658-665.**

**Gusev, Alexander, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen et al. "Integrative approaches for large-scale transcriptome-wide association studies." *Nature genetics* 48, no. 3 (2016): 245-252.**

**Hill, W. David, Neil M. Davies, Stuart J. Ritchie, Nathan G. Skene, Julien Bryois, Steven Bell, Emanuele Di Angelantonio et al. "Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income." *Nature communications* 10, no. 1 (2019): 5741.**

**Pruim, Randall J., Ryan P. Welch, Serena Sanna, Tanya M. Teslovich, Peter S. Chines, Terry P. Gliedt, Michael Boehnke, Gonçalo R. Abecasis, and Cristen J. Willer. "LocusZoom: regional visualization of genome-wide association scan results." *Bioinformatics* 26, no. 18 (2010): 2336-2337.**