Using GW-SEM to analyze latent variables in a multivariate GWAS framework

Brad Verhulst & Joshua Pritikin

March 4th, 2020

TEXAS A&M UNIVERSITY
College of Medicine

VIRGINIA INSTITUTE FOR PSYCHIATRIC AND BEHAVIORAL GENETICS

The Long Road to Boulder

Behav Genet (2017) 47:345–359
DOI 10.1007/s10519-017-9842-6

ORIGINAL RESEARCH

GW-SEM: A Statistical Package to Conduct Genome-Wide Structural Equation Modeling

Brad Verhulst[1] · **Hermine H. Maes**[1] · **Michael C. Neale**[1]

Goals of the Presentation

- Showcase how alternative phenotypic model specifications provide can unique insights into genetic mechanisms
  - Present a latent variable model depicting substance use frequency

- In the practical we are going to conduct an analysis

Strengths of GW-SEM

- More appropriate modeling of the items
  - Ordinal vs Continuous
- Incorporating more environmental and situational variables into the model.
- Move beyond latent variable models into networks, mixture distributions, and multiple group models

Utility of Structural Equation Modeling

Structural Equation Modeling (SEM):

- Incredibly broad framework to estimate a wide variety of statistical models:
  - **Factor Analysis**
  - Path Analysis (mediation, feedback loops)
  - Mixture modeling
  - **Regression (linear, logistic, ordinal)**

Utility of Factor Analysis

Psychometric or Measurement Model:

- The implicit or explicit relationship between the latent variable and its indicators.

- Understanding the measurement of the latent variable will provide valuable insights into the meaning of the factor and interpretation of subsequent associations

Factor Analysis & GWAS

By merging factor analysis with GWAS we can integrate measurement aspects of the phenotype (i.e. the measurement model) into a GWAS.

This allows us to test an enormous range of hypotheses that cannot be tested within a strict linear modeling framework

This can also be used to:

- Assess the fit of the correlations
- assess phenotypic heterogeneity across samples (or subsamples) with measurement invariance

The Benefits of using Raw Data

Raw data gives users more control over the specification of the model components compared with summary statistics from disparate GWAS

Takes a longer to run because we are fitting a separate structural equation model for each SNP

Unparalleled Access to Raw GWAS Data

biobank uk

db GaP
genotypes and phenotypes

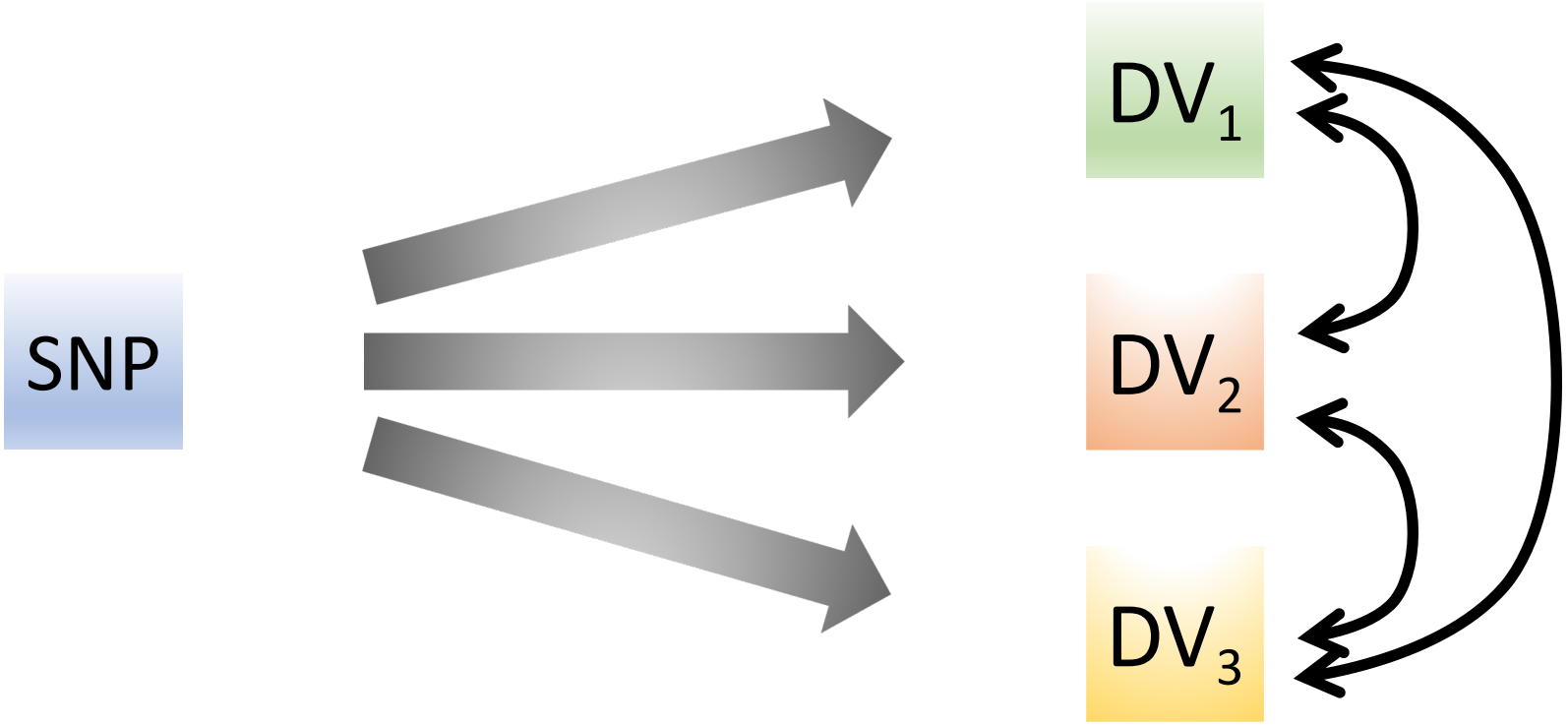
Adolescent Brain Cognitive Development®
Teen Brains. Today's Science. Brighter Future.

All of Us
RESEARCH PROGRAM

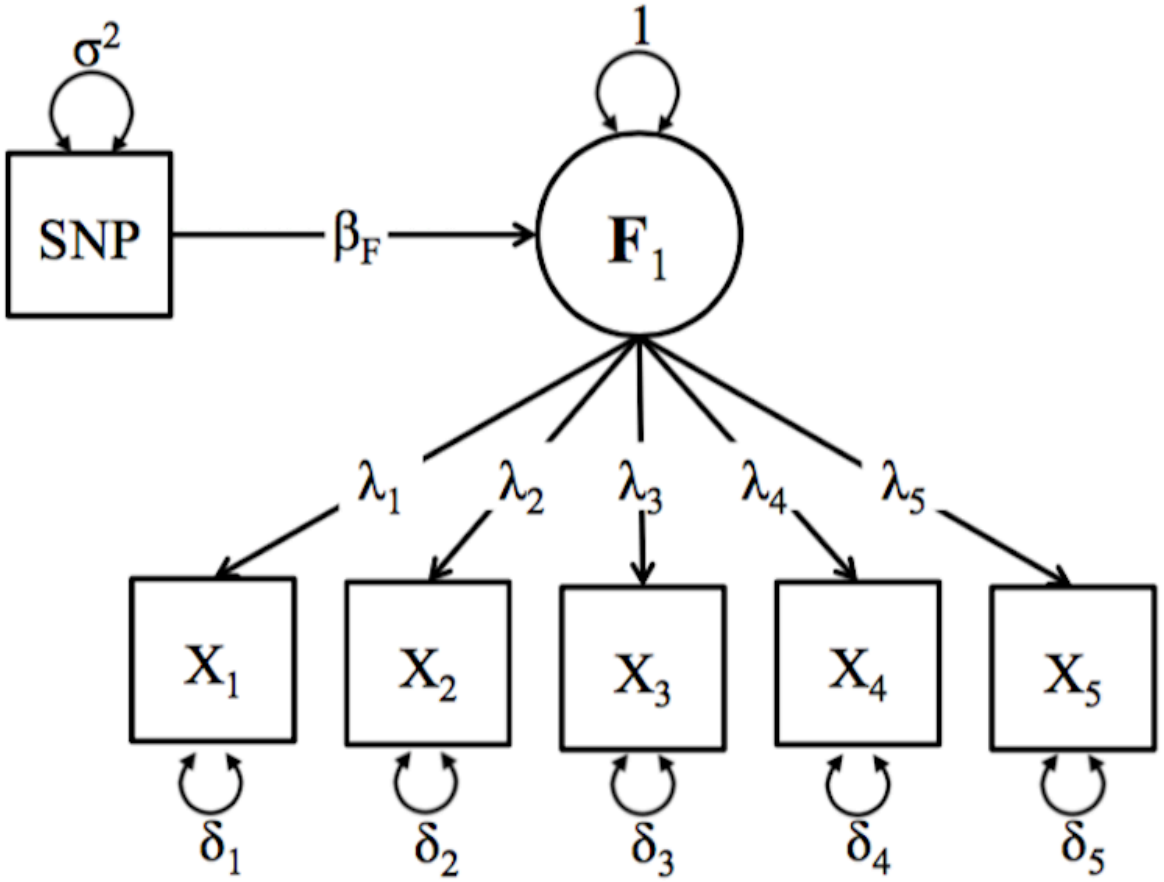
Standard GWAS Model

SNP → DV

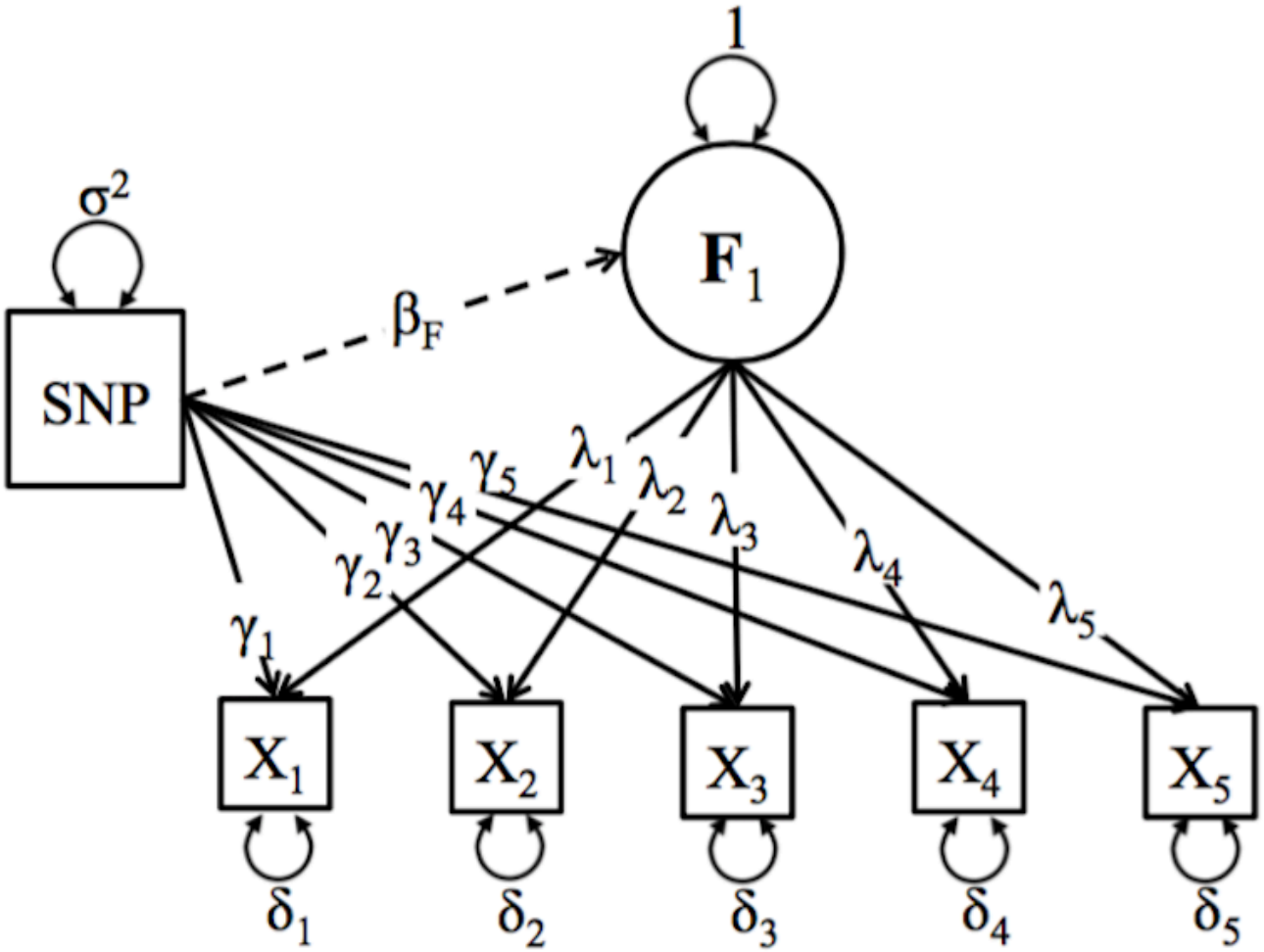
Novel Hypothesis Tests



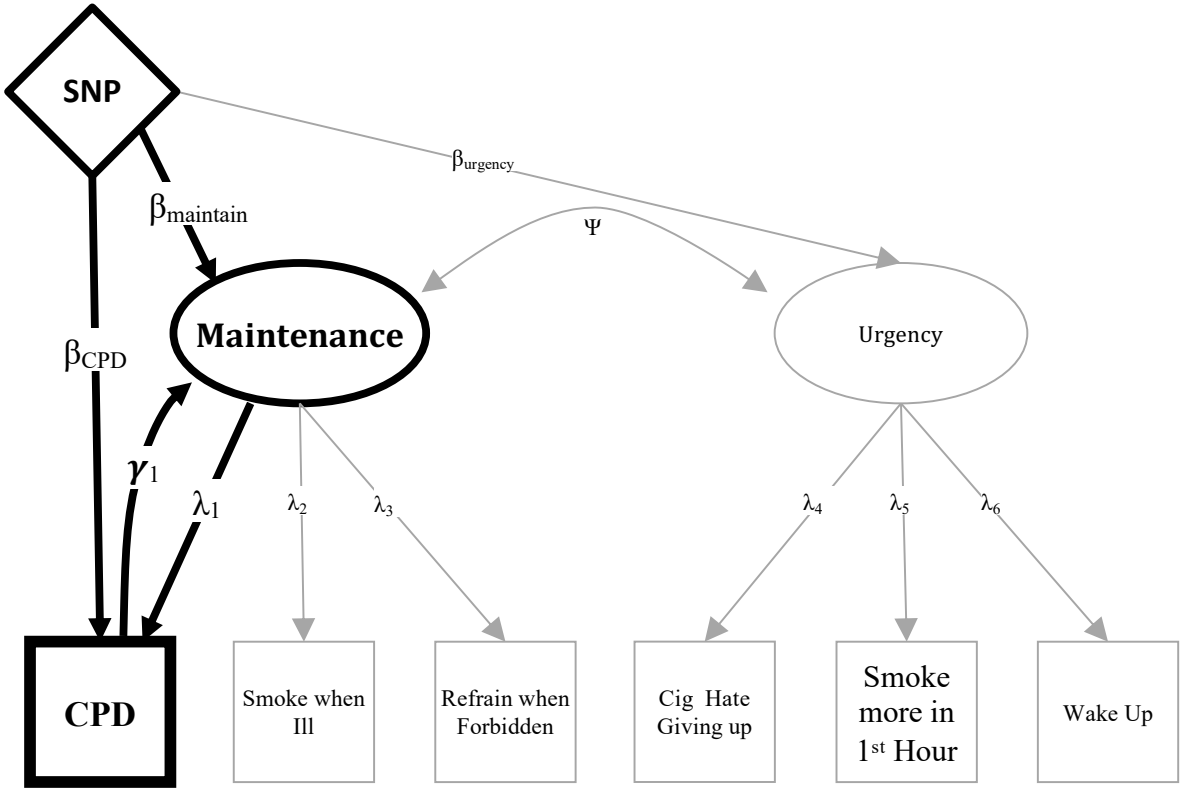
Novel Hypothesis Tests



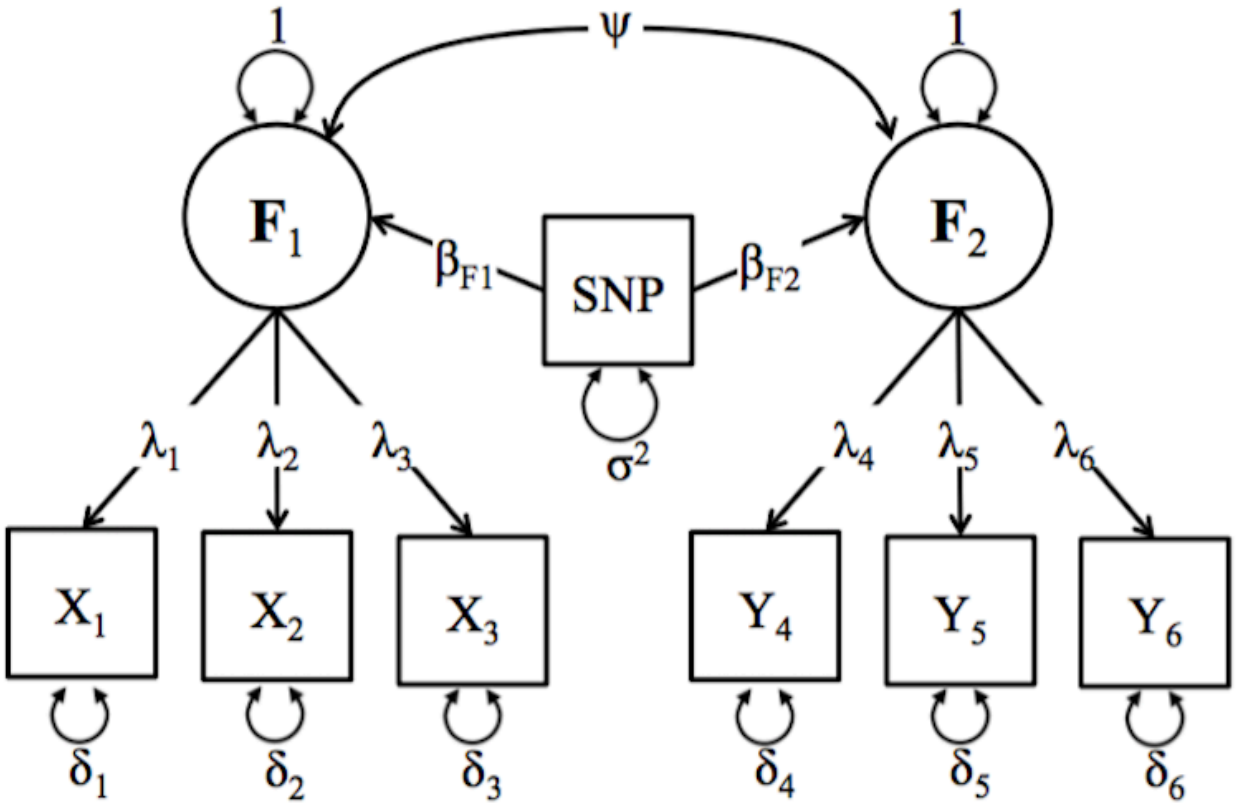
Novel Hypothesis Tests



FTND Example



Novel Hypothesis Tests



Real Data Example

Method

- Data were obtained from the UK Biobank (Application 40967)
  - Individuals included in the analysis if they were :
    - of European ancestry
    - unrelated to other individuals (one person selected from families)
    - had sufficient genotyping quality

- Total Sample Size for the analysis was 379, 153
  - a subsample was asked about cannabis use (N=112, 109)

- Fit a single factor confirmatory factor model to frequency of use for tobacco, cannabis and alcohol (not quantity)
  1. SNP predicted the latent factor
  2. SNP predicted the individual item residuals

Question Wording and Response Options

Tobacco

In the past, how often have you smoked tobacco?

Smoked on most or all days
Smoked occasionally
Just tried once or twice
I have never smoked

N = 348, 032

Cannabis

Have you taken CANNABIS (marijuana, grass, hash, ganja, blow, draw, skunk, weed, spliff, dope), even if it was a long time ago?

Yes, more than 100 times
Yes, 11-100 times
Yes, 3-10 times
Yes, 1-2 times
No

N = 112, 109

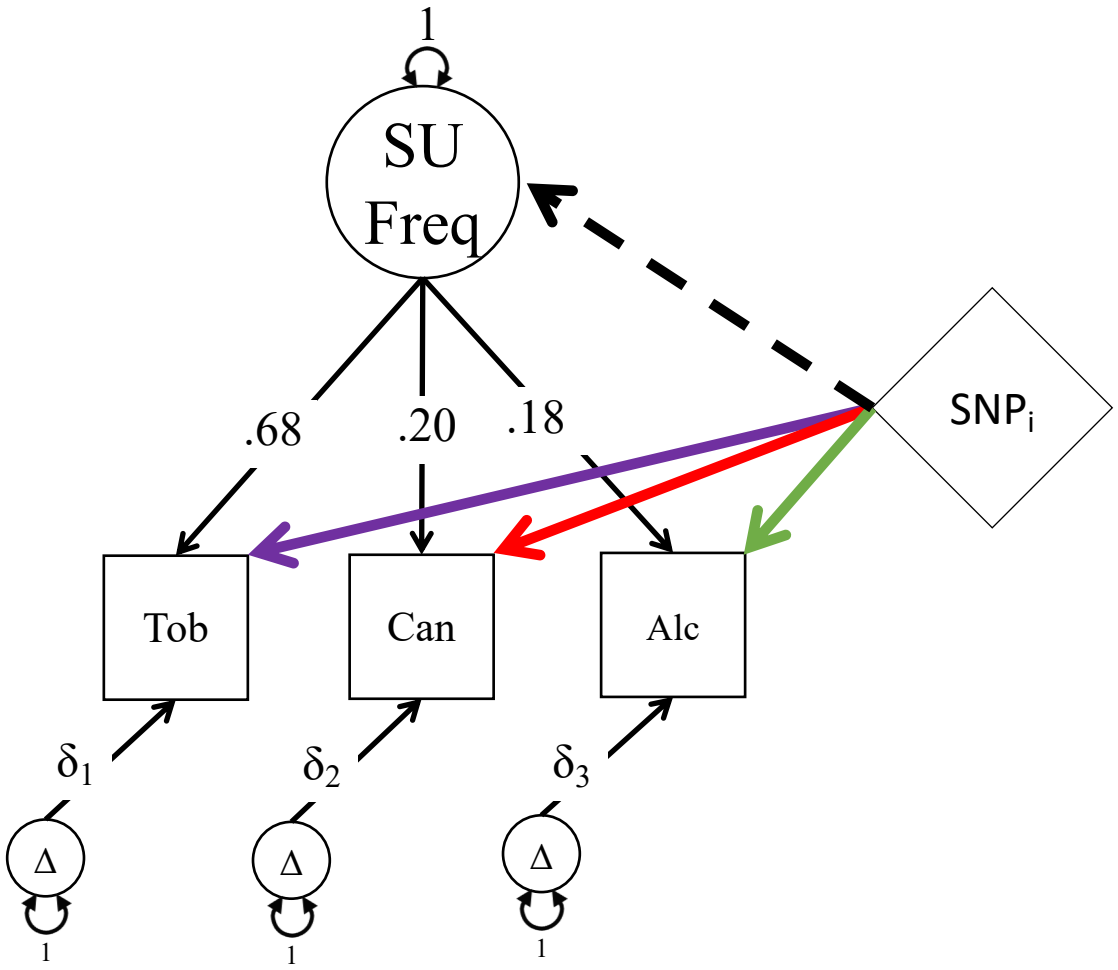
Alcohol

About how often do you drink alcohol?

Daily or almost daily
Three or four times a week
Once or twice a week
One to three times a month
Special occasions only
Never

N = 379, 153

Single Factor Model of Substance Use Frequency



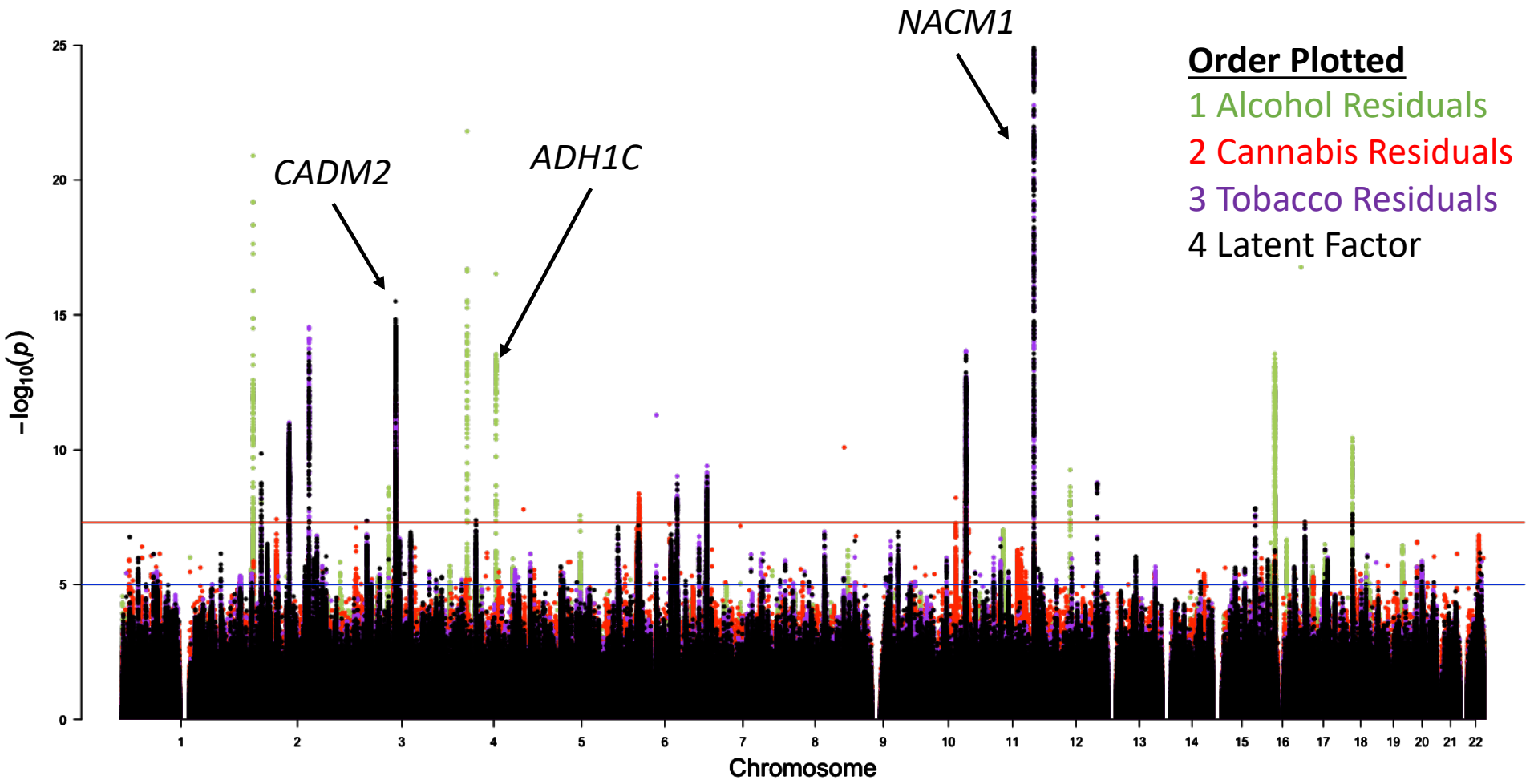
The latent SU Freq variable is disproportionately driven by tobacco frequency.

This means the regression of the latent factor on the SNP will resemble the tobacco frequency residual regression more than the other variables

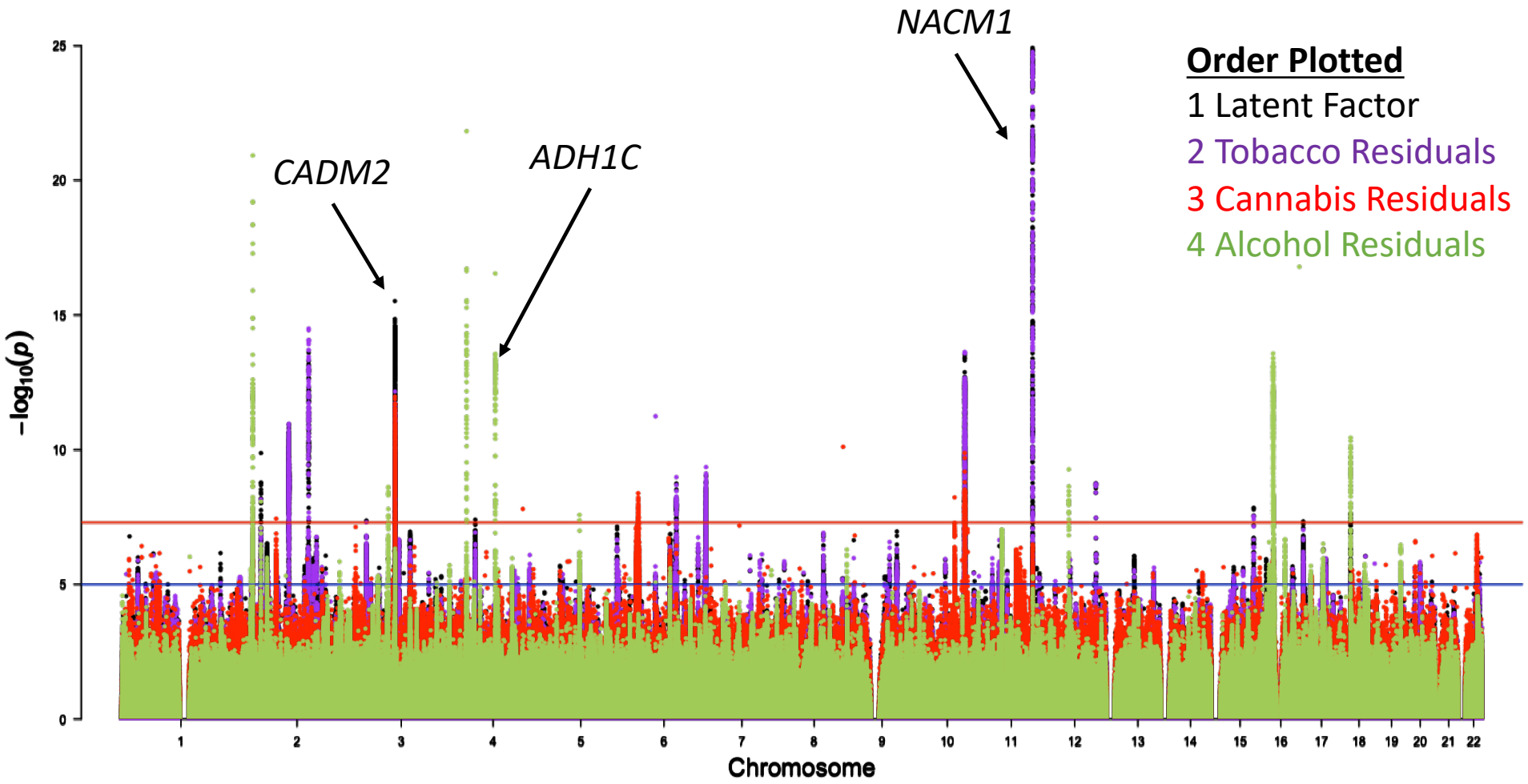
While model fit is excellent, this is not a very good model from a psychometric perspective

Not Shown: All models controlled for Age, Sex, and the top 10 ancestry PCs

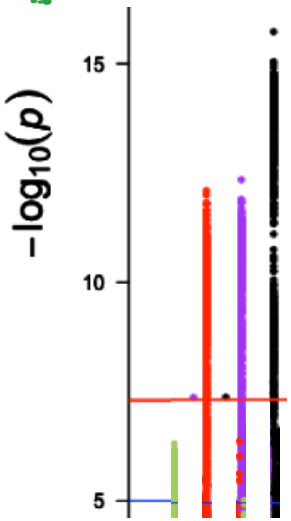
Layered Manhattan Plot of the Associations with the Latent Factor and Items Residuals

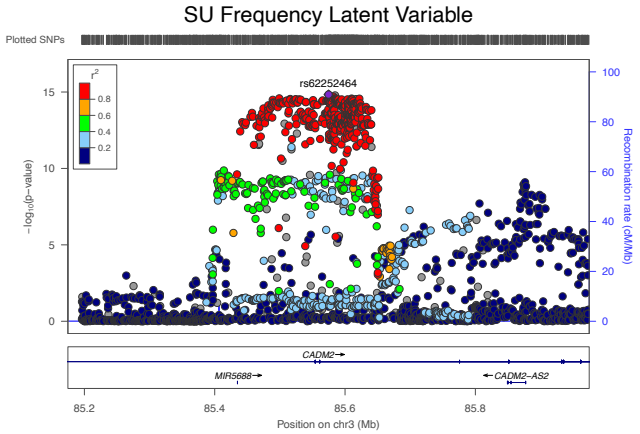


Layered Manhattan Plot of the Associations with the Latent Factor and Items Residuals

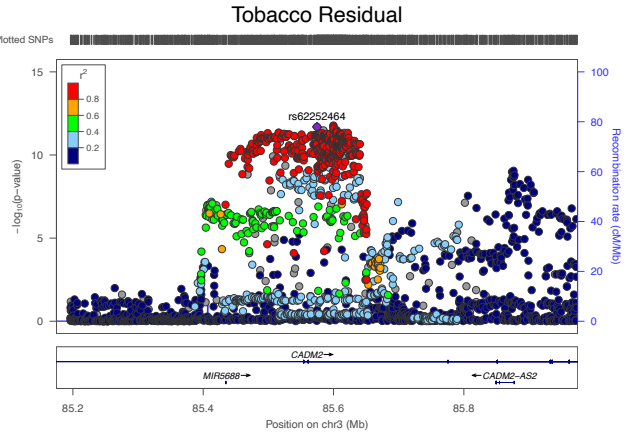


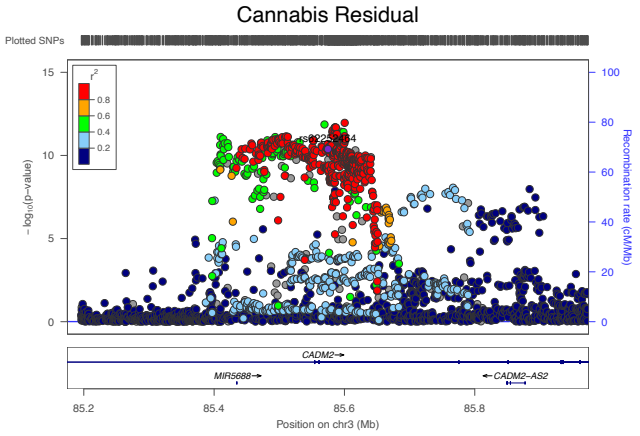
Locus Zoom Plot of *CADM2*

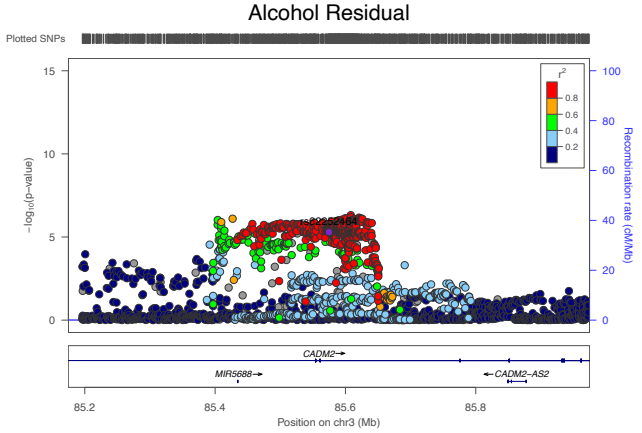




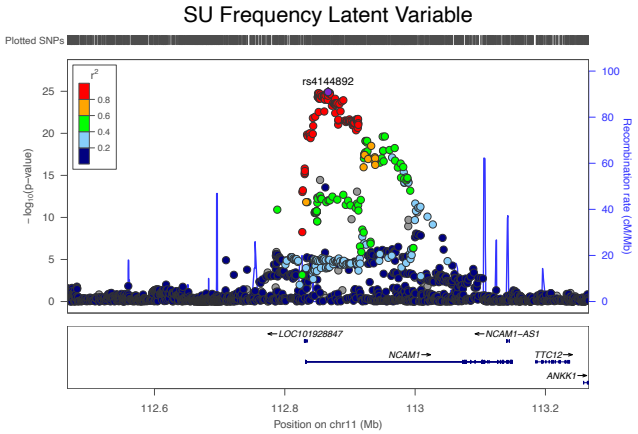
The combination of multiple frequency items increases the signal for the latent variable



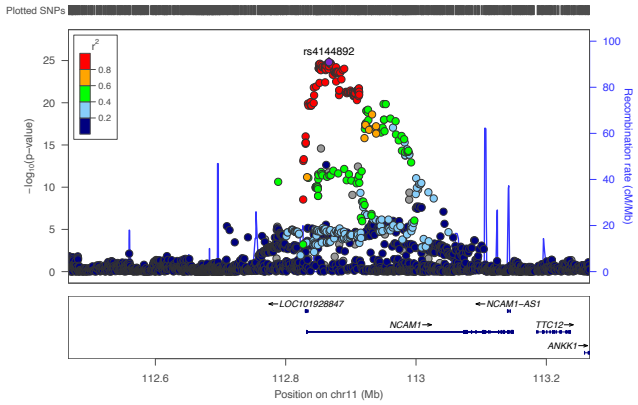


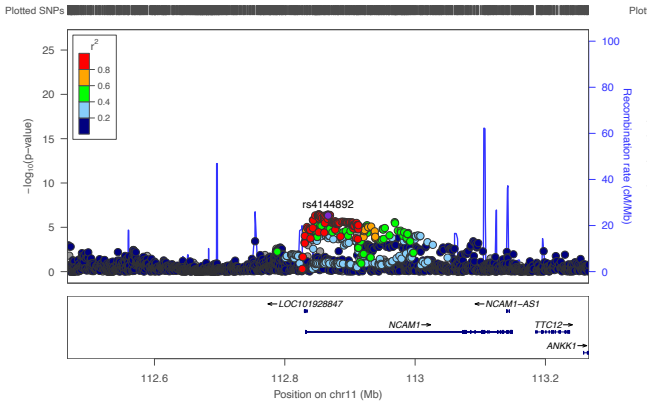


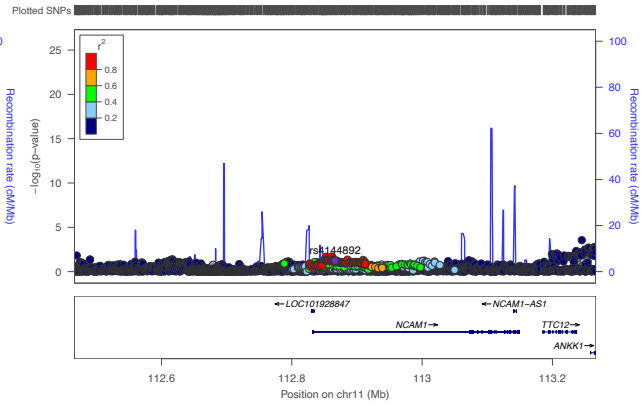
Locus Zoom Plot of *NCAM1*



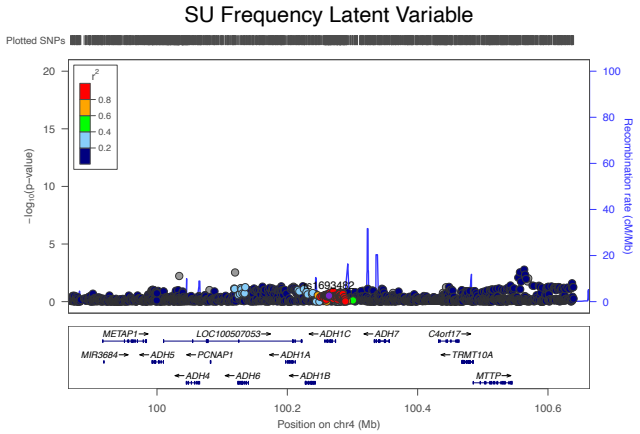
The latent variable signal is coming almost exclusively from tobacco



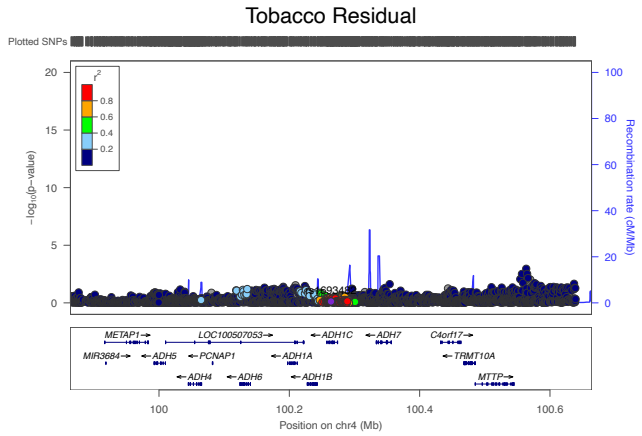


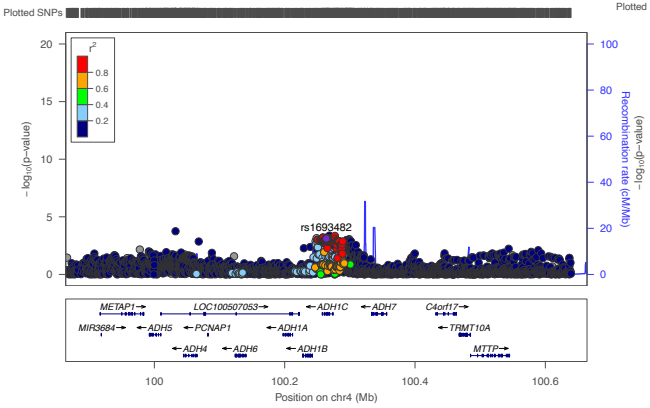


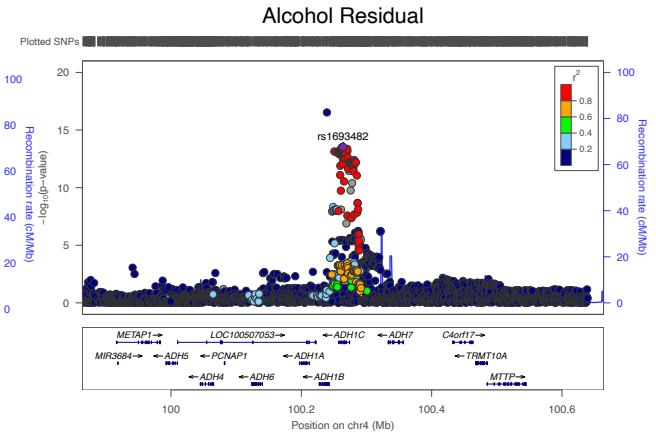
Locus Zoom Plot of *ADH1C*



There is not genomic signal for the latent variable, but a strong signal from drinking frequency







Three Types of Associations

1. More signal from the latent factor than the items
   - Ideal use of a factor model to enhance our GWAS of the latent trait
   - CADM2 has previously been associated with risk taking behaviors
2. Association with a single item that dominates that factor
   - NCAM1 seemed less like a SU frequency association and more like a smoking related association
3. Association with a residual item but not the latent factor
   - This indicates substance specificity
   - ADH1C is not really a SU related gene, but instead is an alcohol related gene

Leveraging Latent Variable Models

- SEM allows us to integrate the phenotypic measurement of constructs into the genomic
  - Let the genes “revise” the interpretation of the constructs

- Integrating rarely used substances into factor models
  - Potential to infer loci that are associated with general use even if:
    - sample sizes are small
    - specific substance use behavior is not measured at all (potentially)
  - Proxy GWAS for rarely used substances

- We can analyze competitive responses that can teach us more about tradeoffs between substances than would could understand by any other type of modeling framework.

Post GWAS Software Integration

Easy integration with:

- LD Score Regression Software
  - LDhub
  - genomicSEM
- LocusZoom
- qqman

Working on integrating with:

- MAGMA, FUMA and other Post-analysis software

Example Syntax

```
require(gwsem)

addData <- as.data.frame(read.table("addict2.txt", header = T))
addData$sex <- as.numeric(addData$sex)
addData$age <- as.numeric(addData$age)/10

addFac <- buildOneFac(phenoData = addData, itemNames  = c("nic", "can", "alc"),
                      covariates = c("sex", "age", "pc1", "pc2", "pc3", "pc4",
                                     "pc5", "pc6",  "pc7" ,"pc8","pc9","pc10"),
                      fitfun = "WLS")

GWAS(depFac, snpData= "../ukb_01/chr1.pgen", out = "addFacChr1.log", SNP = 1:100000)
```

Acknowledgements

Texas A&M Collaborator

Shaunna Clark

VCU Collaborators

Elizabeth Prom-Wormley

Michael Neale

This project used data from the UK Biobank that was accessed under application number 40967