# Genome-Wide Structural Equation Modeling

Joshua N. Pritikin & Brad Verhulst

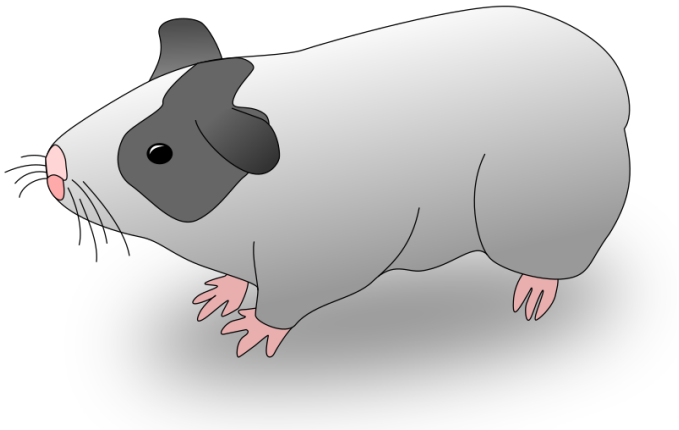
Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University

Feb 2020

# Acknowledgment

Mentors and collaborators

- Mike Neale
- Rob Kirkpatrick
- OpenMx development team



Funded in part by NIDA R25 DA-26119

# Single-nucleotide polymorphism

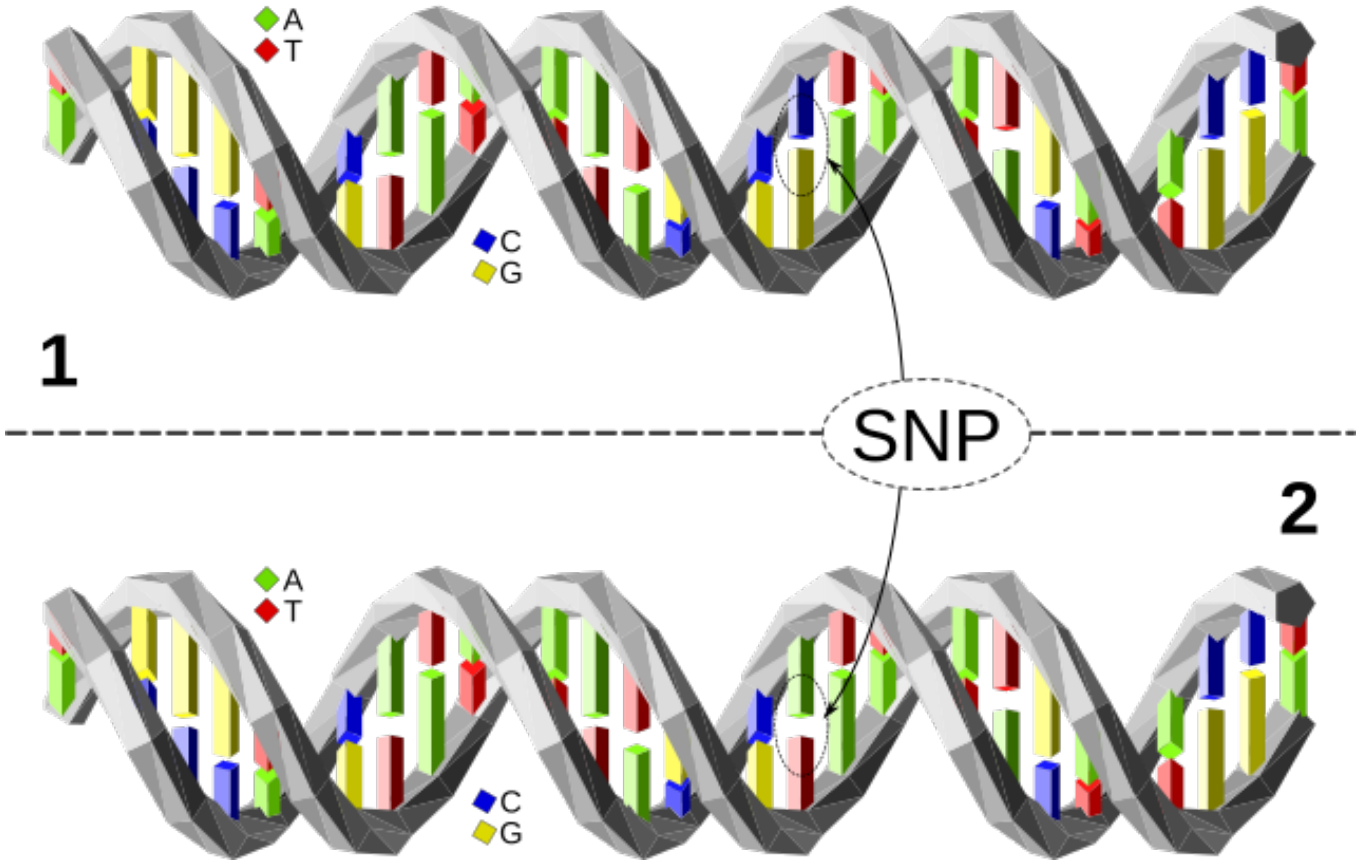
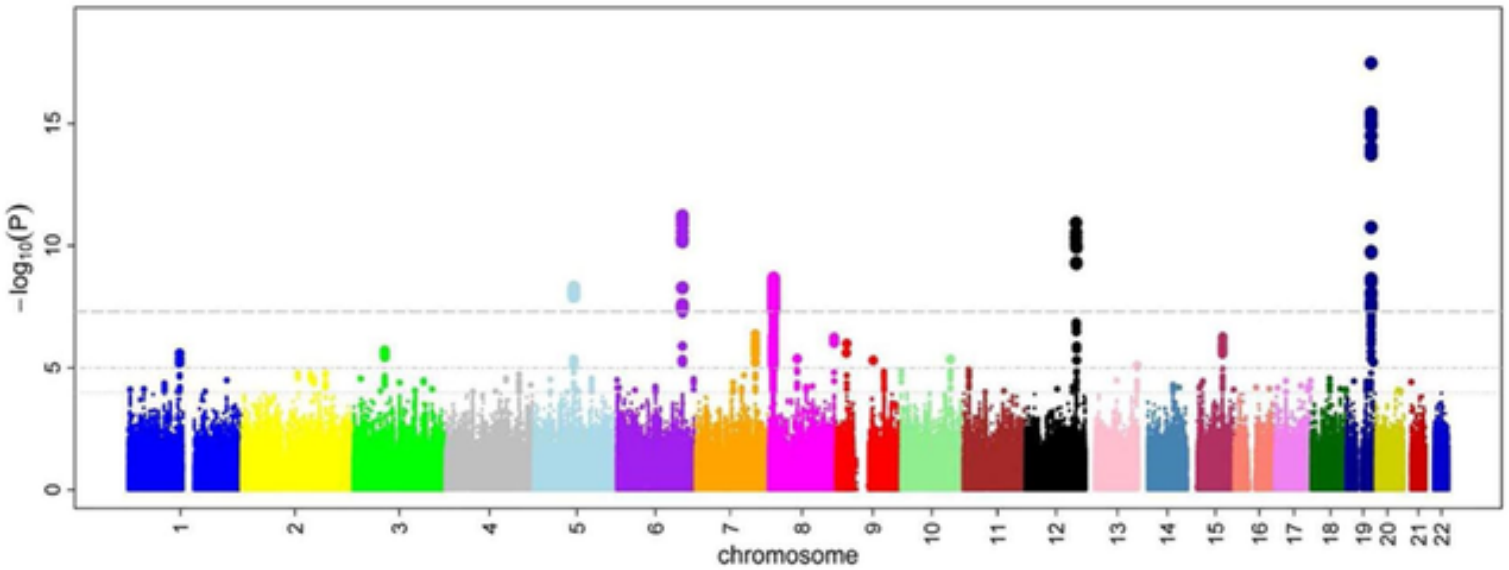



Image by David Eccles (gringer), CC BY 4.0,
`https://commons.wikimedia.org/w/index.php?curid=2355125`

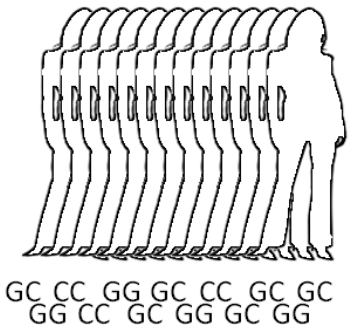
# Genome-wide association studies




Ordinary regressions of SNP on microcirculation[1]

[1] By M. Kamran Ikram et al (2010) PLoS Genet. Oct 28;6(10):e1001184. CC BY 2.5, `https://commons.wikimedia.org/w/index.php?curid=18056138`

# Case control design



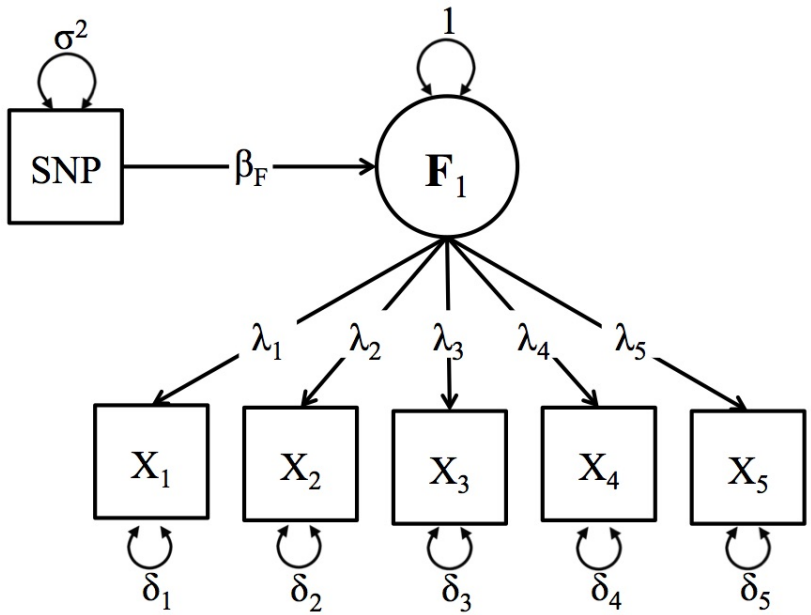

| | *SNP1* | *SNP2* | *SNP ...* |
|---|---|---|---|
| | **Cases** | **Cases** | *Repeat for all SNPs* |
| | Count of G: 2104 of 4000 | Count of G: 1648 of 4000 | |
| | Frequency of G: 52.6% | Frequency of G: 41.2% | |
| | **Controls** | **Controls** | |
| | Count of G: 2676 of 6000 | Count of G: 2532 of 6000 | |
| | Frequency of G: 44.6% | Frequency of G: 42.2% | |
| | **P-value:** $5.0 \cdot 10^{-15}$ | **P-value:** 0.33 | |

Probit or logit regression[2]

[2] Lasse Folkersen CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=18062562

# Other models?

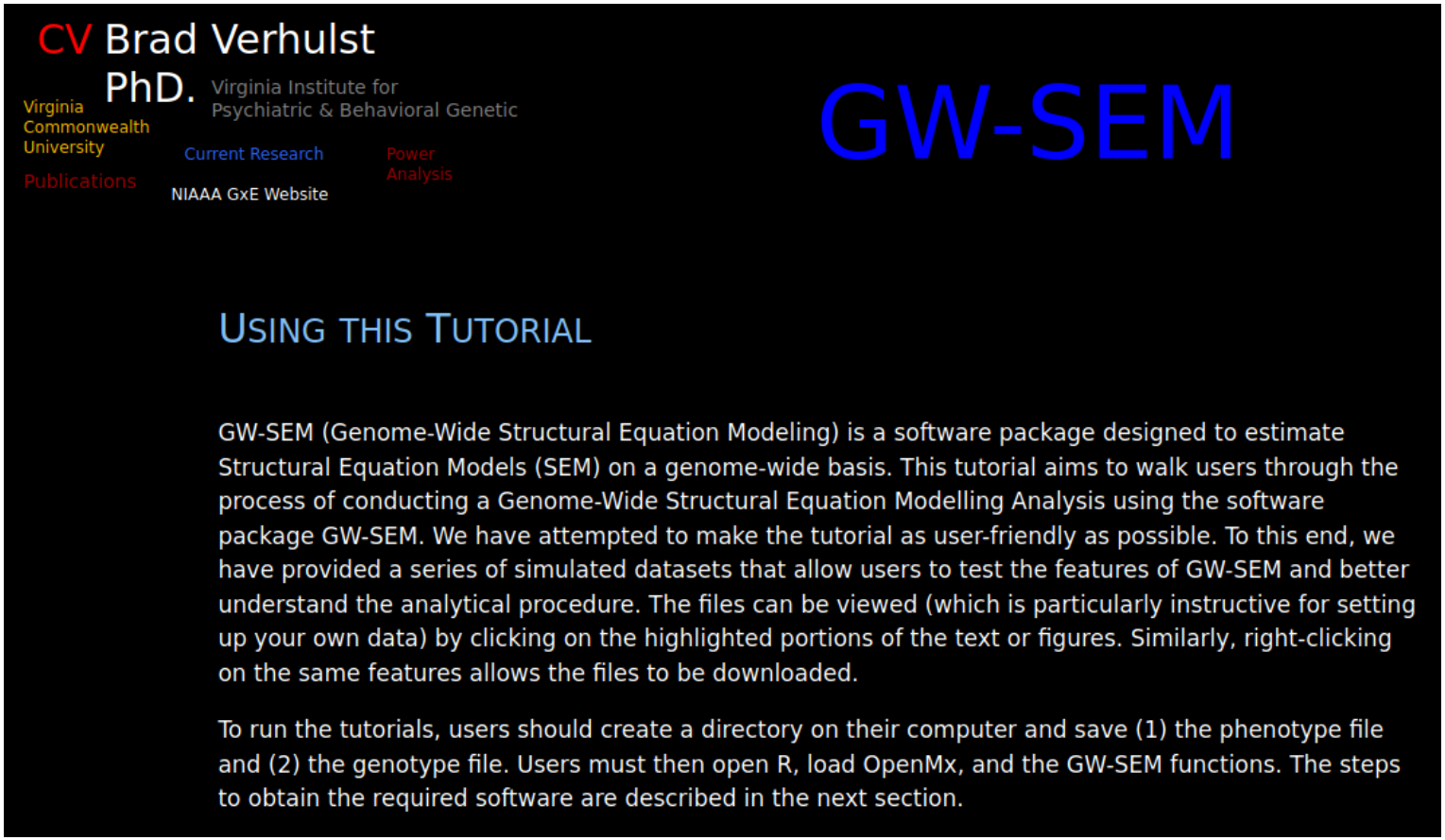
Regression,[3] but what about complex phenotypes?




With millions of SNPs, the main challenge is performance

[3] Hirschhorn and Daly (2005)

# Initial version in 2017




# Retrospective 1/2

Advances[4]

- Factor models
- Choice of weighted least squares (WLS) or ML

[4] Verhulst, Maes, and Neale (2017)

# Retrospective 2/2

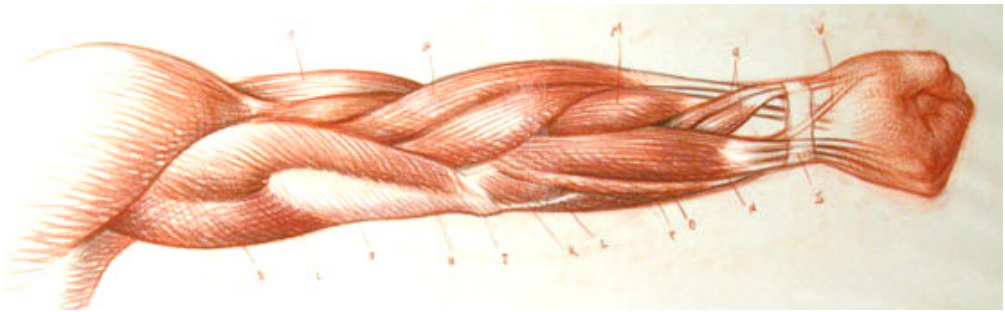
Deficiencies

- Barely adequate performance
- Not on CRAN
- Awkward user interface
- Time consuming, format conversion step
- Incorrect standard errors
- Limited to a few models

# Tour

```
library(gwsem)   # on CRAN                                    1
```

# Anatomy of the API



```
model <- buildOneFac(phenodata, paste0('i', 1:5))                    1
GWAS(model, 'example.pgen','out.log')                                2
got <- loadResults('out.log', 'snp_to_F',                            3
                   signAdj='lambda_i1')                              4
plot(got)  # Manhattan                                               5
```

Talking points

- `model` is an OpenMx `MxModel`
- `GWAS` fits the model against every SNP in `example.pgen`
- `loadResults` computes P-values for the focal parameter

# Predefined models

Model construction

- `buildItem` – regression, but can do multiple items
- `buildOneFac` – single factor model
  similar to GEMMA & plink MANOVA
- `buildOneFacRes` – single factor residuals model
- `buildTwoFac` – two factor model (pleiotropy, comorbidity)

Continuous or ordinal indicators

# GWAS function


Read

- bgen (UK BioBank)
- pgen and bed (plink)

using

- High performance C++
- Uncompress in streaming mode

# Results

`isSuspicious` – heuristic to classify SNP results

- `loadResults`
- `loadSuspicious`

# Substance use application

# 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 379153
  - a subsample was asked about cannabis use (N=112109)

- 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 = 379153

## 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 = 112109

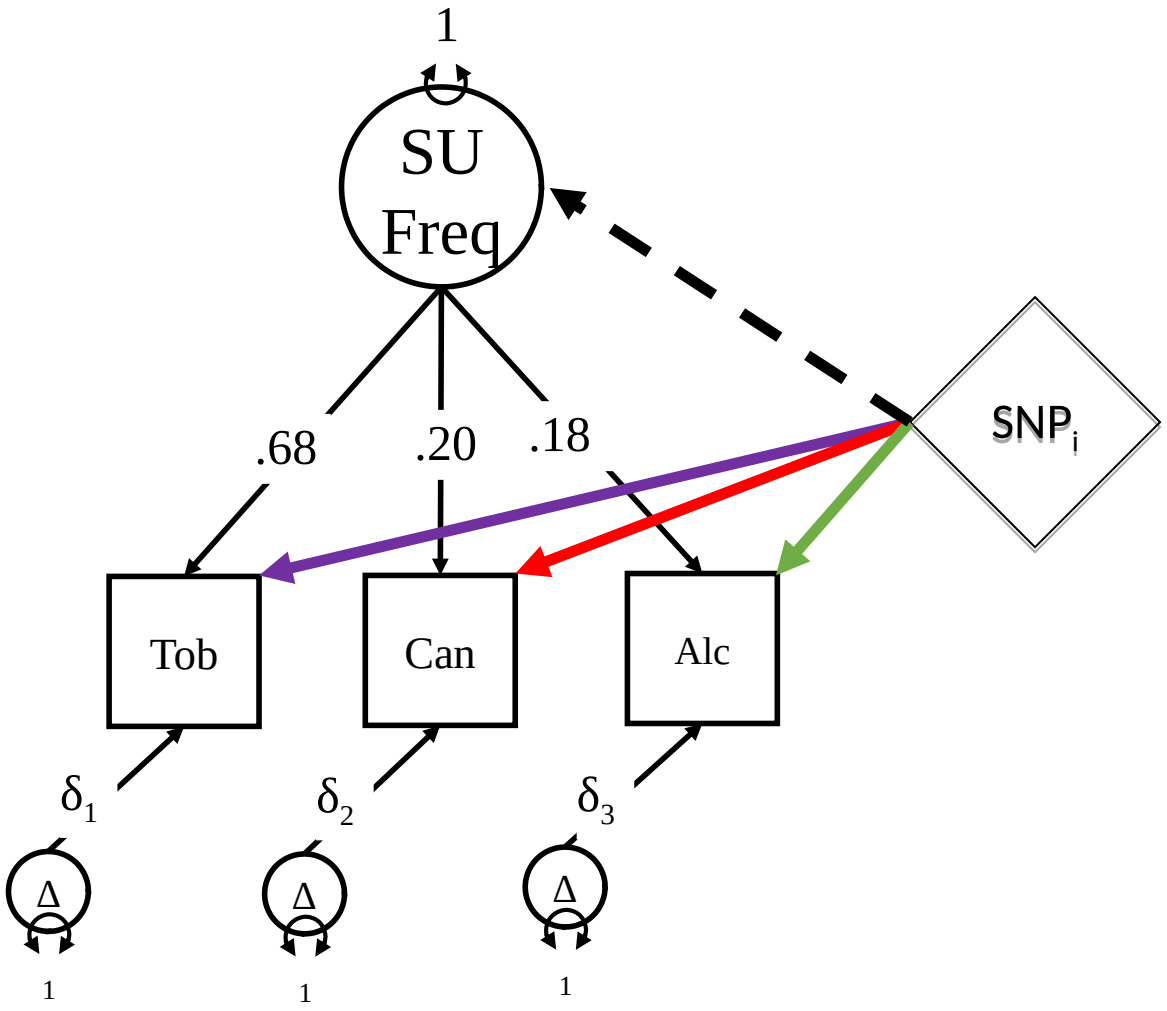
## Alcohol

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 = 379153

# Single Factor Model of Substance Use Frequency



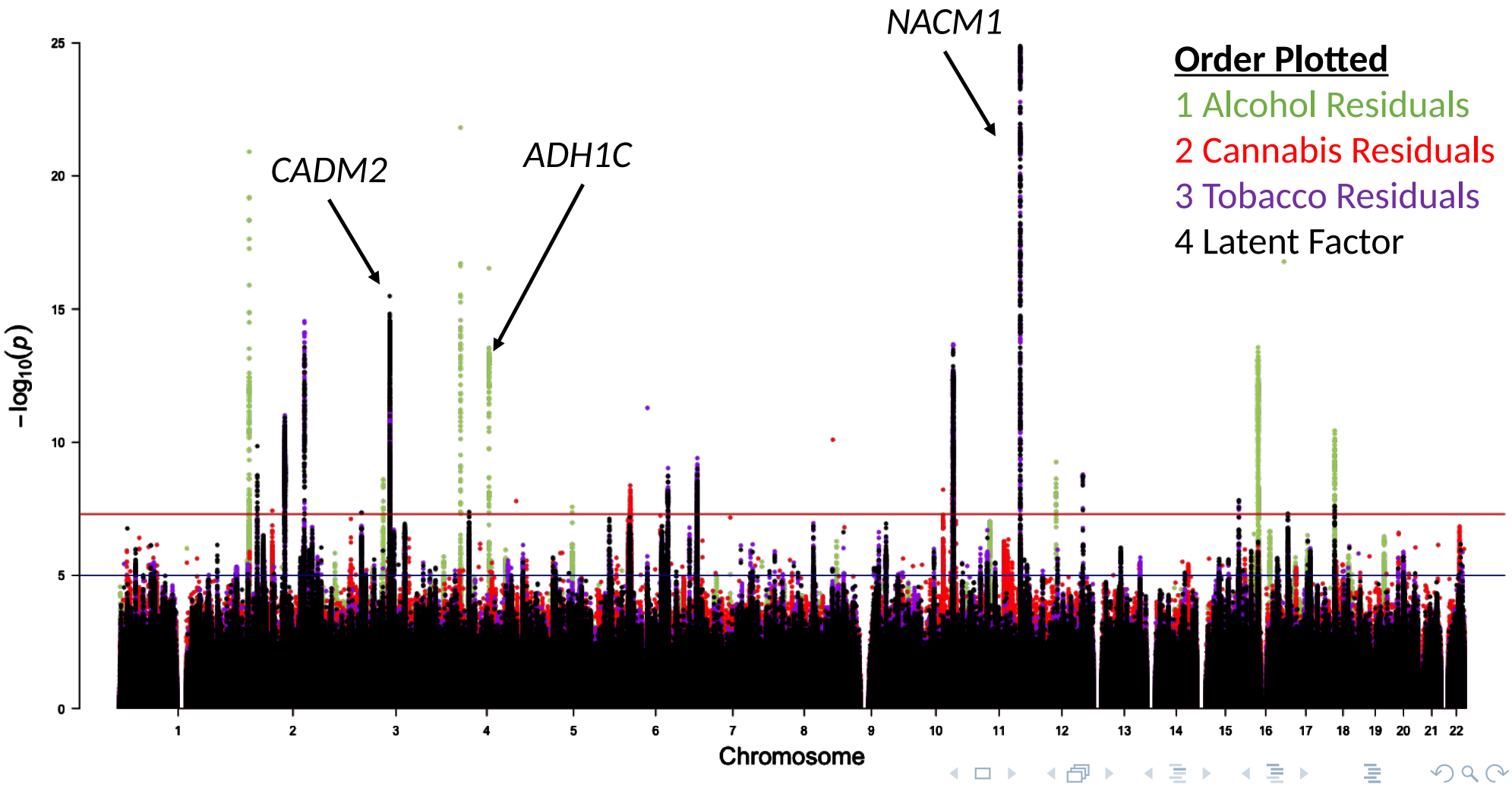

The latent SU Freq variable is disproportionately driven by tobacco frequency.

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

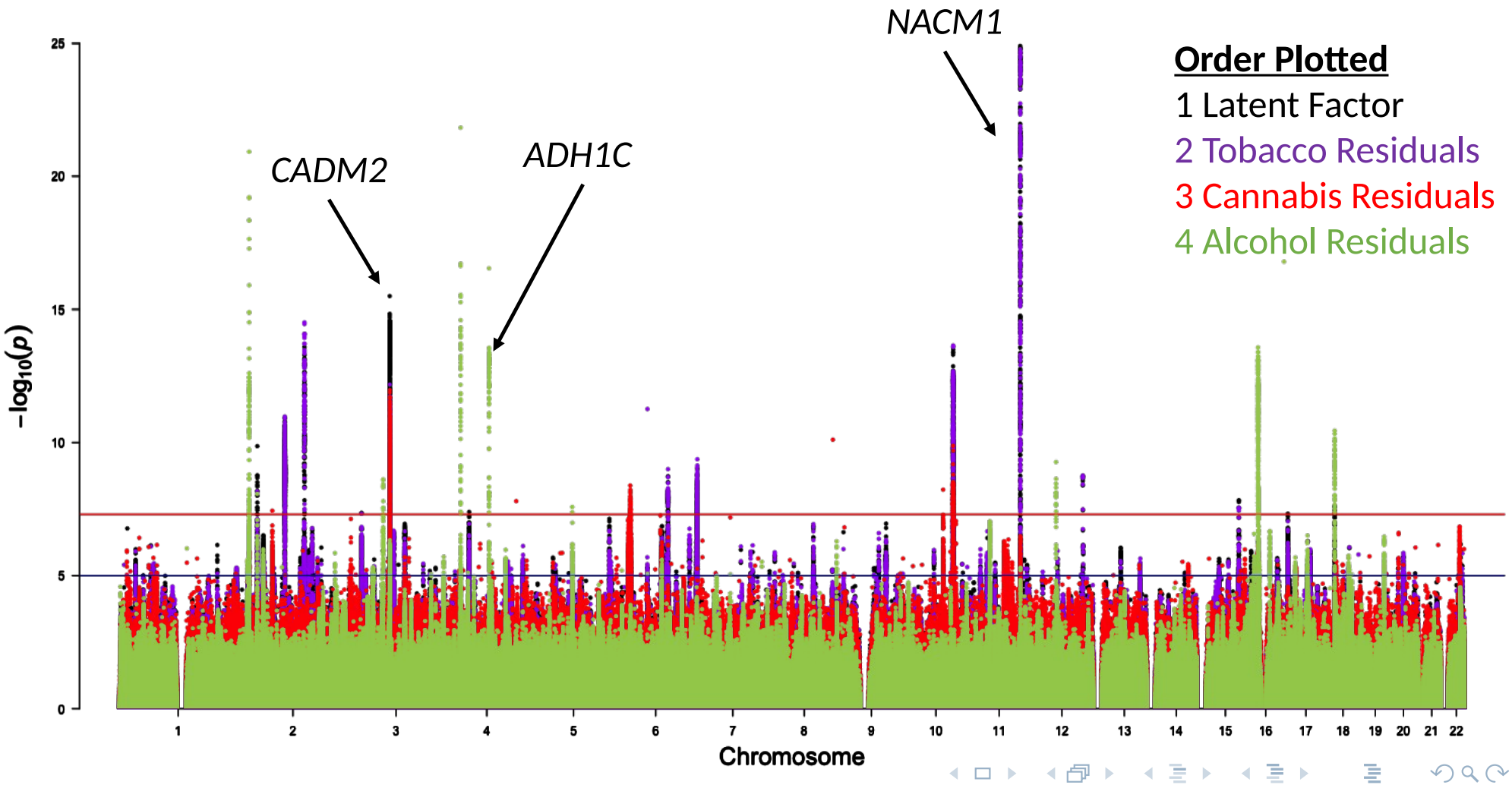
While model fit is excellent, this is not a very good model from a psychometric perspective

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

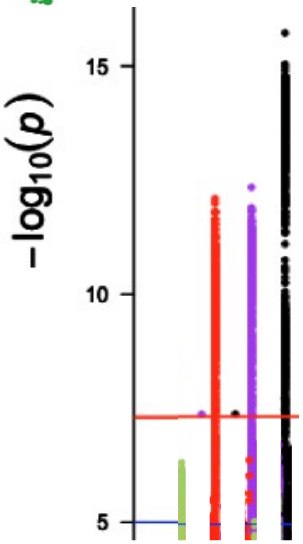
# Layered Manhattan Plot of the Associations with the Latent Factor and Items Residuals

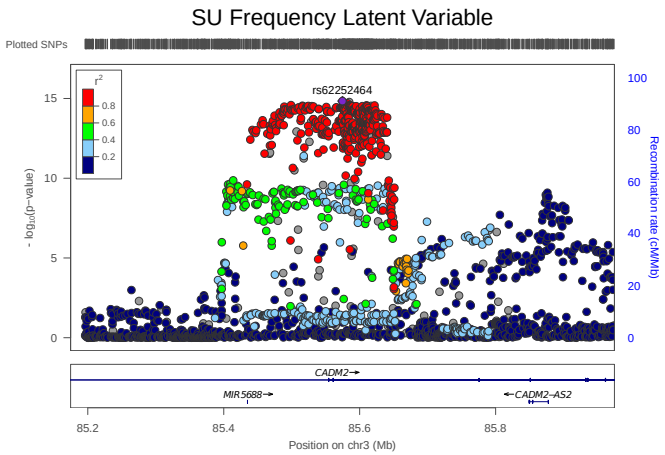


# Layered Manhattan Plot of the Associations with the Latent Factor and Items Residuals

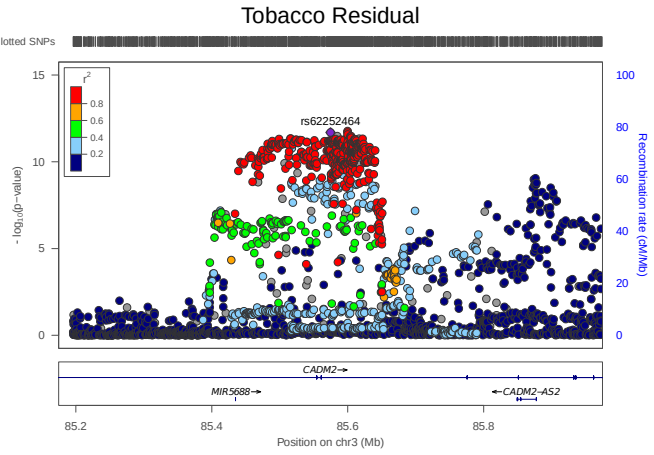


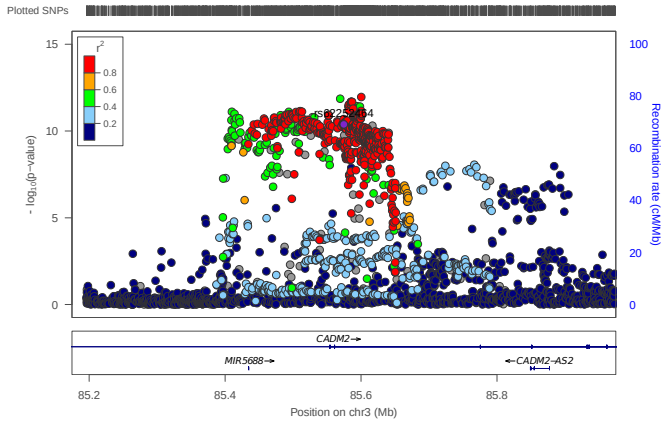
# Locus Zoom Plot of *CADM2*

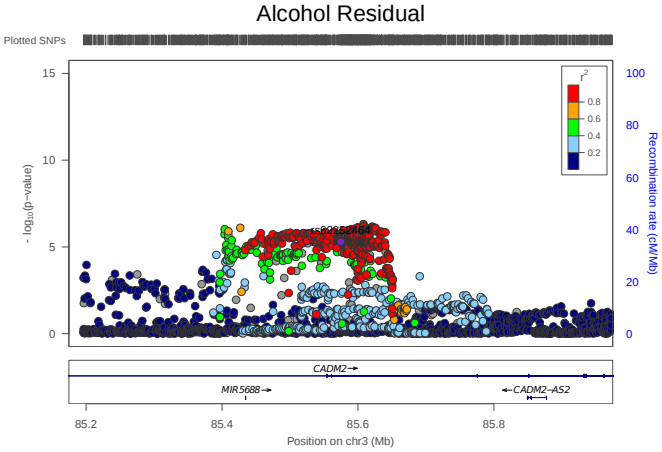


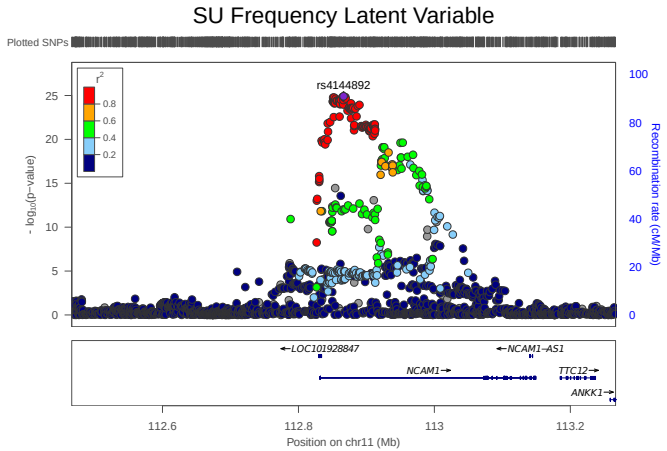

The combination of multiple frequency items increases the signal for the latent variable



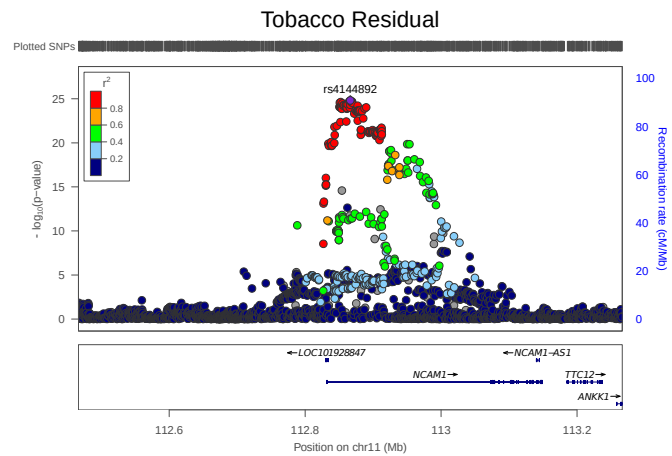



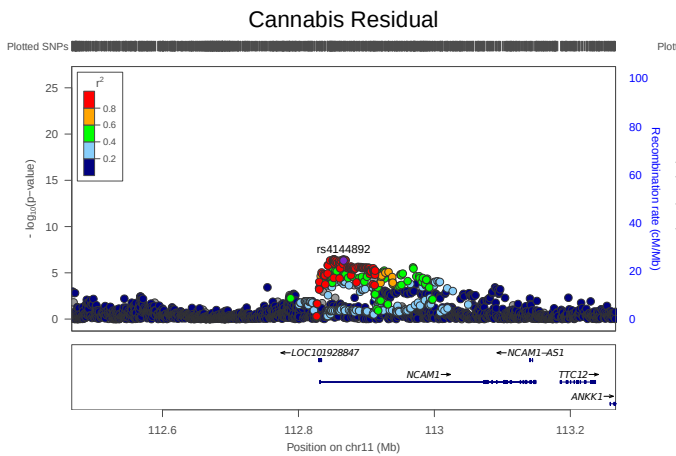



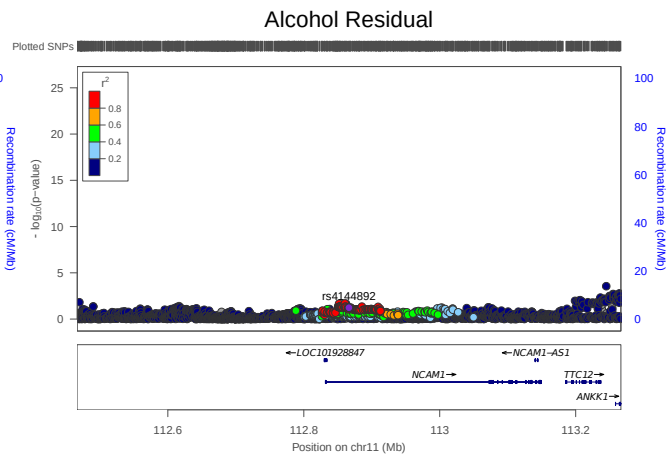
# Locus Zoom Plot of *NCAM1*



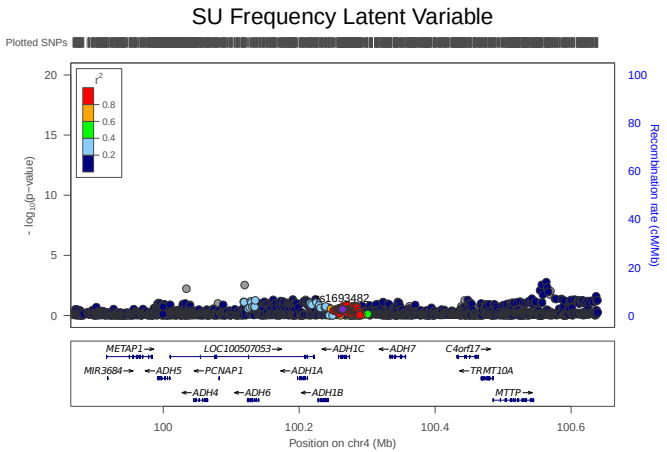

The latent variable signal is coming almost exclusively from tobacco



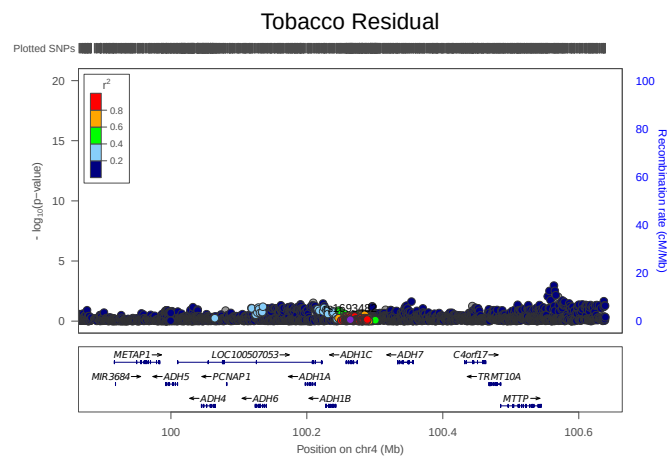



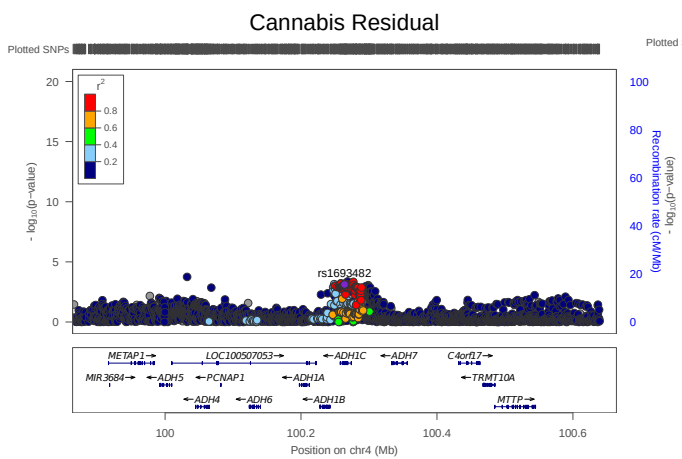



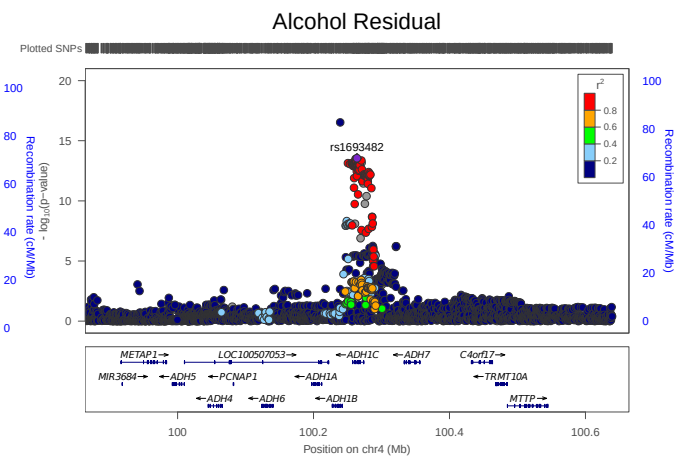
# Locus Zoom Plot of *ADH1C*




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
   - NACM1 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 SU 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

# Retrospective 2/2

Deficiencies

- Barely adequate performance
- Not on CRAN
- Awkward user interface
- Time consuming, format conversion step
- Incorrect standard errors
- Limited to a few models

# Current status

Advances

- Barely adequate performance → comparable to similar software
- Not on CRAN → On CRAN
- Awkward user interface → much improved
- Time consuming, format conversion step → bgen, plink support
- Incorrect standard errors → now correct
- Limited to a few models → arbitrary OpenMx models
  (workflow: Onyx → OpenMx → gwsem)

# Future

Coming soon

- Gene environment interactions
- Mixture models, network models, quantity initiation models
- Distmix integration[5]
- Comparison to summary GWAS analyses (e.g. Genomic SEM[6])

[5] Lee et al. (2015)
[6] Grotzinger et al. (2019)

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., ... others (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*, *3*(5), 513–525. doi: 10.1038/s41562-019-0566-x

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95–108. doi: 10.1038/nrg1521

Lee, D., Bigdeli, T. B., Williamson, V. S., Vladimirov, V. I., Riley, B. P., Fanous, A. H., & Bacanu, S.-A. (2015, 06). DISTMIX: Direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. *Bioinformatics*, *31*(19), 3099-3104. doi: 10.1093/bioinformatics/btv348

Verhulst, B., Maes, H. H., & Neale, M. C. (2017). GW-SEM: A statistical package to conduct genome-wide structural equation modeling. *Behavior Genetics*, *47*(3), 345–359. doi: 10.1007/s10519-017-9842-6