Genomic SEM

David Evans[1,2,3*]

1 Institute for Molecular Bioscience, University of Queensland

2 University of Queensland Diamantina Institute

3 MRC Integrative Epidemiology Unit, University of Bristol

*with thanks Ben Neale, Michel Nivard, Andrew Grotzinger

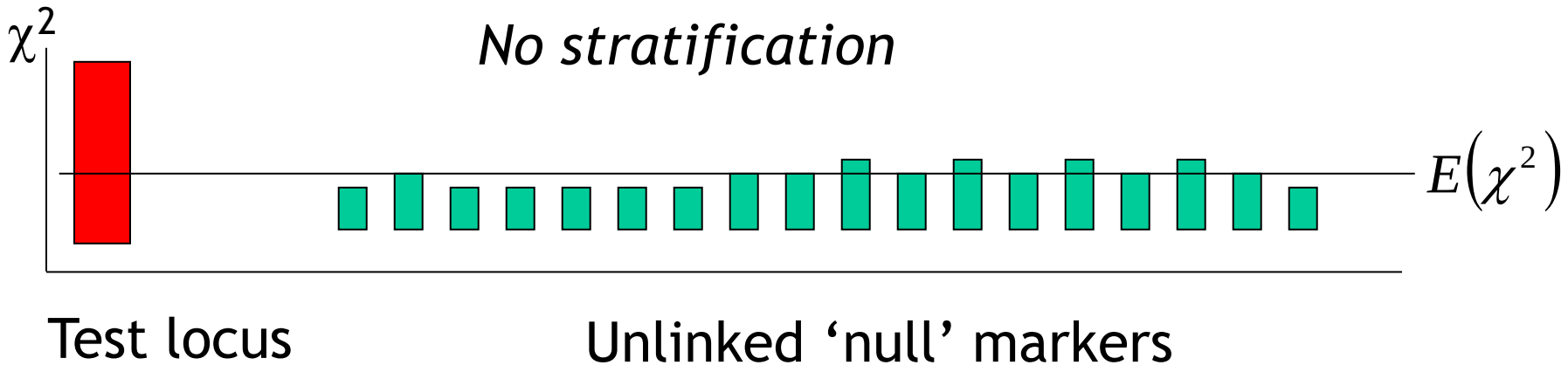
What is Genomic SEM?

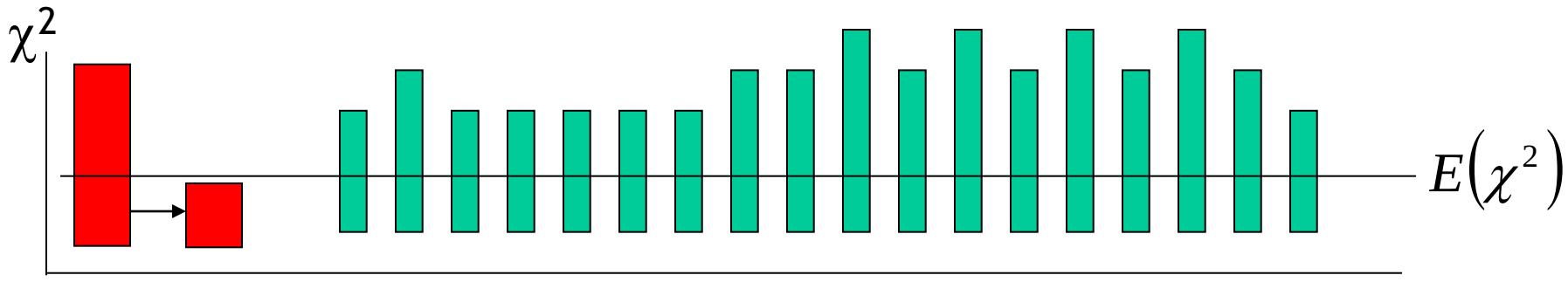
- Similar to ordinary SEM but uses a genetic variance-covariance matrix rather than a phenotypic covariance matrix

- The genetic variance-covariance matrix is usually derived from the analysis of genome-wide summary statistics data (rather than individual level genetic and phenotypic data)

- It uses a different fit function to traditional SEM (Diagonally weighted least squares)

Genomic control





Stratification $\rightarrow$ *adjust test statistic*

Genomic inflation factor and Genomic Control

- "λ" is Genome-wide inflation factor

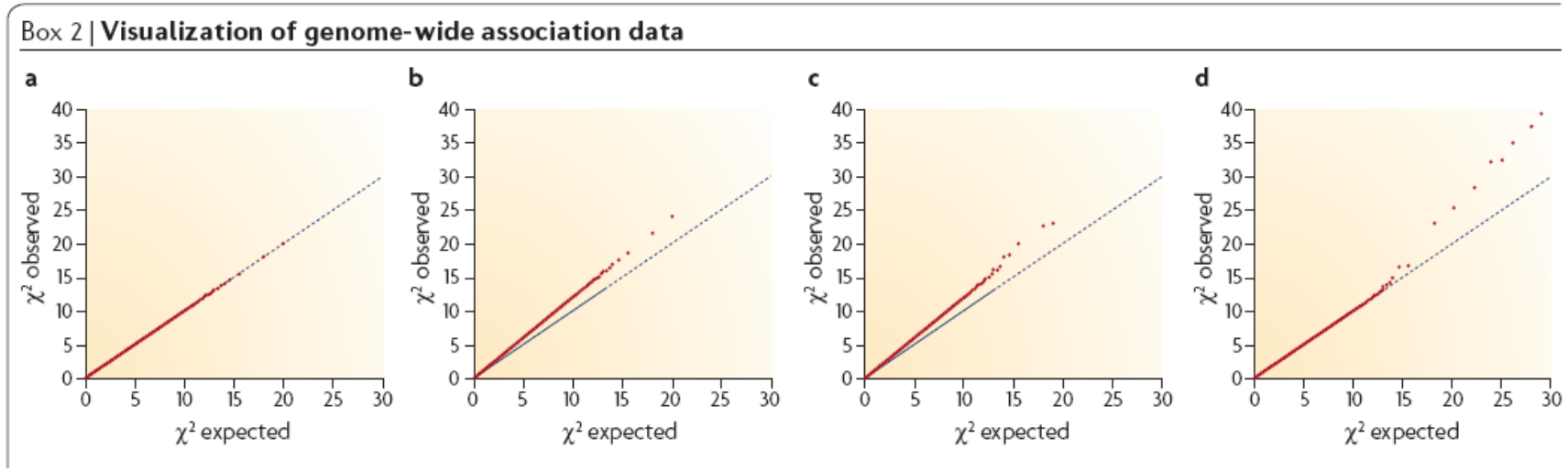
$$\hat{\lambda} = median\{\chi_1^2, \chi_2^2, \ldots, \chi_N^2\} / 0.455$$

- Test statistic is distributed under the null:

$$T_N / \lambda \sim \chi^2{}_1$$

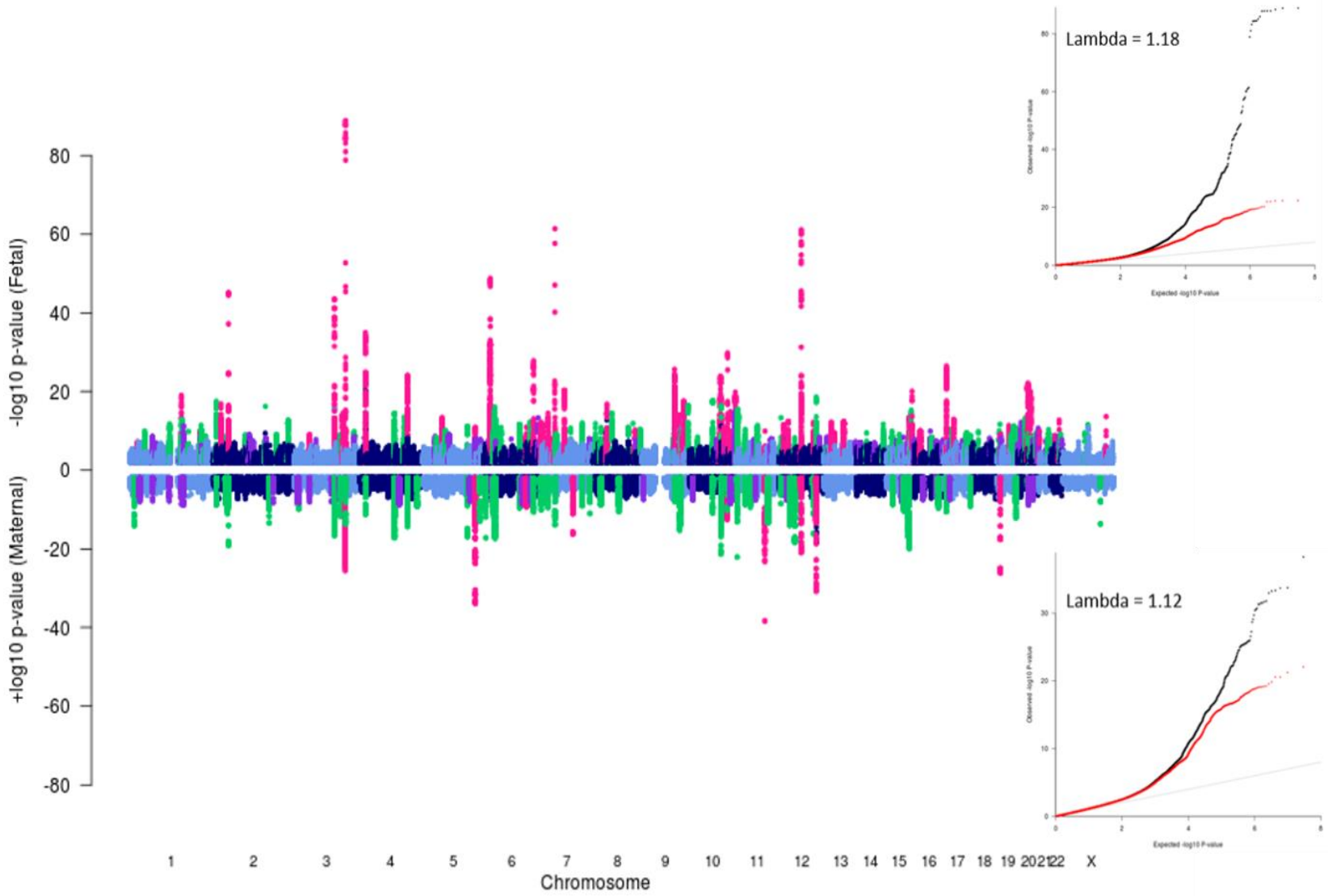
- Problems…

QQ plots



McCarthy et al. (2008) Nature Genetics

Polygenicity vs Type 1 Error



Warrington et al. (2019) *Nat Genet*

LD Score Regression

nature genetics

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan[1–3], Po-Ru Loh[1,4], Hilary K Finucane[4,5], Stephan Ripke[2,3], Jian Yang[6], Schizophrenia Working Group of the Psychiatric Genomics Consortium[7], Nick Patterson[1], Mark J Daly[1–3], Alkes L Price[1,4,8] & Benjamin M Neale[1–3]

Both polygenicity (many small genetic effects) and confounding biases, such as cryptic relatedness and population stratification, can yield an inflated distribution of test statistics in genome-wide association studies (GWAS). However, current methods cannot distinguish between inflation from a true polygenic signal and bias. We have developed an approach, LD Score regression, that quantifies the contribution of each by examining the relationship between test statistics and linkage disequilibrium (LD). The LD Score regression intercept can be used to estimate a more powerful and accurate correction factor than genomic control. We find strong evidence that polygenicity accounts for the majority of the inflation in test statistics in many GWAS of large sample size.

Variants in LD with a causal variant show an elevation in test statistics in association analysis proportional to their LD (measured by r^2) with the causal variant[1–3]. The more genetic variation an index variant tags, the higher the probability that this index variant will tag a causal variant. In contrast, inflation from cryptic relatedness within or between cohorts[4–6] or population stratification purely from genetic drift will not correlate with LD.

Under a polygenic model, in which effect sizes for variants are drawn independently from distributions with variance proportional to $1/(p(1-p))$, where p is the minor allele frequency (MAF), the expected χ^2 statistic of variant j is:

$$E[\chi^2 \mid \ell_j] = Nh^2\ell_j / M + Na + 1 \quad (1)$$

where N is the sample size; M is the number of SNPs, such that h^2/M is the average heritability explained per SNP; a measures the contribution of confounding biases, such as cryptic relatedness and population stratification; and $\ell_j = \Sigma_k r^2_{jk}$ is the LD Score of variant j, which measures the amount of genetic variation tagged by j (a full derivation of this equation is provided in the **Supplementary Note**). This relationship holds for meta-analyses and also for ascertained studies of binary phenotypes, in which case h^2 is on the observed scale. Consequently, if we regress the χ^2 statistics from GWAS against LD Score (LD Score regression), the intercept minus one is an estimator of the mean contribution of confounding bias to the inflation in the test statistics.

RESULTS

Overview of methods

We estimated LD Scores from the European-ancestry samples in the 1000 Genomes Project[7] (EUR) using an unbiased estimator[8] of r^2 with 1-cM windows, singletons excluded (MAF > 0.13%) and no r^2 cutoff. Standard errors were estimated by jackknifing over blocks of individuals, and we used these standard errors to correct for attenuation bias in LD Score regression (that is, the downward bias in the magnitude of the regression slope that occurs when the regressor is measured noisily; Online Methods).

For LD Score regression, we excluded variants with EUR MAF < 1% because the LD Score standard errors for these variants were very high (note that the variants included in LD Score regression are a subset of the variants included in LD Score estimation). In addition, we excluded loci with extremely large effect sizes or extensive long-range LD from all regressions because these loci can be considered outliers in such an analysis and would have disproportionate influence on the regression (Online Methods).

An important consideration in the estimation of LD Score is the extent to which the sample from which LD Score is estimated matches the sample for the association study. If there is a mismatch between the LD Scores from the reference population and the target population used for GWAS, then LD Score regression can be biased in two ways. First, if LD Scores in the reference population are equal to LD Scores in the target population plus mean-zero noise, then the intercept will

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [5]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [7]A full list of members and affiliations appears in the **Supplementary Note**. [8]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. Correspondence should be addressed to B.M.N. (bneale@broadinstitute.org).

Received 7 March 2014; accepted 7 January 2015; published online 2 February 2015; doi:10.1038/ng.3211

nature genetics

An atlas of genetic correlations across human diseases and traits

Brendan Bulik-Sullivan[1–3,9], Hilary K Finucane[4,9], Verneri Anttila[1–3], Alexander Gusev[5,6], Felix R Day[7], Po-Ru Loh[1,5], ReproGen Consortium[8], Psychiatric Genomics Consortium[8], Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3[8], Laramie Duncan[1–3], John R B Perry[7], Nick Patterson[1], Elise B Robinson[1–3], Mark J Daly[1–3], Alkes L Price[1,5,6,10] & Benjamin M Neale[1–3,10]

Identifying genetic correlations between complex traits and diseases can provide useful etiological insights and help prioritize likely causal relationships. The major challenges preventing estimation of genetic correlation from genome-wide association study (GWAS) data with current methods are the lack of availability of individual-level genotype data and widespread sample overlap among meta-analyses. We circumvent these difficulties by introducing a technique—cross-trait LD Score regression—for estimating genetic correlation that requires only GWAS summary statistics and is not biased by sample overlap. We use this method to estimate 276 genetic correlations among 24 traits. The results include genetic correlations between anorexia nervosa and schizophrenia, anorexia and obesity, and educational attainment and several diseases. These results highlight the power of genome-wide analyses, as there currently are no significantly associated SNPs for anorexia nervosa and only three for educational attainment.

Understanding the complex relationships among human traits and diseases is a fundamental goal of epidemiology. Randomized controlled trials and longitudinal studies are time-consuming and expensive, so many potential risk factors are studied using cross-sectional correlation studies performed for a single time point. Obtaining causal inferences from such studies can be challenging because of issues such as confounding and reverse causation, which can lead to spurious associations and mask the effects of real risk factors[1,2]. Genetics can help elucidate cause and effect, as inherited genetic risks cannot be subject to reverse causation and are correlated with a smaller list of confounders.

The first methods to test for genetic overlap were family studies[3–7]. To estimate the genetic overlap for many pairs of phenotypes, family study designs require the measurement of multiple traits for the same individuals. Consequently, it is challenging to scale these designs to a large number of traits, especially traits that are difficult or costly to measure (for example, low-prevalence diseases). More recently, GWAS have allowed effect size estimates to be obtained for specific genetic variants, so it is possible to test for shared genetics by looking for correlations in effect sizes across traits, which does not require measuring multiple traits per individual.

There exists a large class of methods for interrogating genetic overlap via GWAS that focus only on genome-wide significant SNPs. One of the most influential methods in this class is Mendelian randomization, which uses significantly associated SNPs as instrumental variables to attempt to quantify causal relationships between risk factors and disease[1,2]. Methods that focus on significant SNPs are effective for traits where there are many significant associations that account for a substantial fraction of heritability[8,9]. For many complex traits, heritability is distributed over thousands of variants with small effects, and the proportion of heritability accounted for by significantly associated variants at current sample sizes is small[10]. In such situations, one can often obtain more accurate results by using genome-wide data rather than data for only significantly associated variants[11].

A complementary approach is to estimate genetic correlation, which considers the effects of all SNPs, including those that do not reach genome-wide significance (Online Methods). The two main existing techniques for estimating genetic correlation from GWAS data are restricted maximum likelihood (REML)[11–16] and polygenic scores[17,18]. These methods have only been applied to a few traits because they require individual-level genotype data, which are difficult to obtain owing to informed consent limitations.

To overcome these limitations, we have developed a technique for estimating genetic correlation using only GWAS summary statistics that is not biased by sample overlap. Our method, cross-trait LD Score

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Stanley Center for Psychiatric Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [7]Medical Research Council (MRC) Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [8]A full list of members and affiliations appears in the **Supplementary Note**. [9]These authors contributed equally to this work. [10]These authors jointly supervised this work. Correspondence should be addressed to B.B.-S. (bulik@broadinstitute.org), B.M.N. (bneale@broadinstitute.org), H.K.F. (hilaryf@mit.edu) or A.L.P. (aprice@hsph.harvard.edu).

Received 2 February; accepted 26 August; published online 28 September 2015; doi:10.1038/ng.3406

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, 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 & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

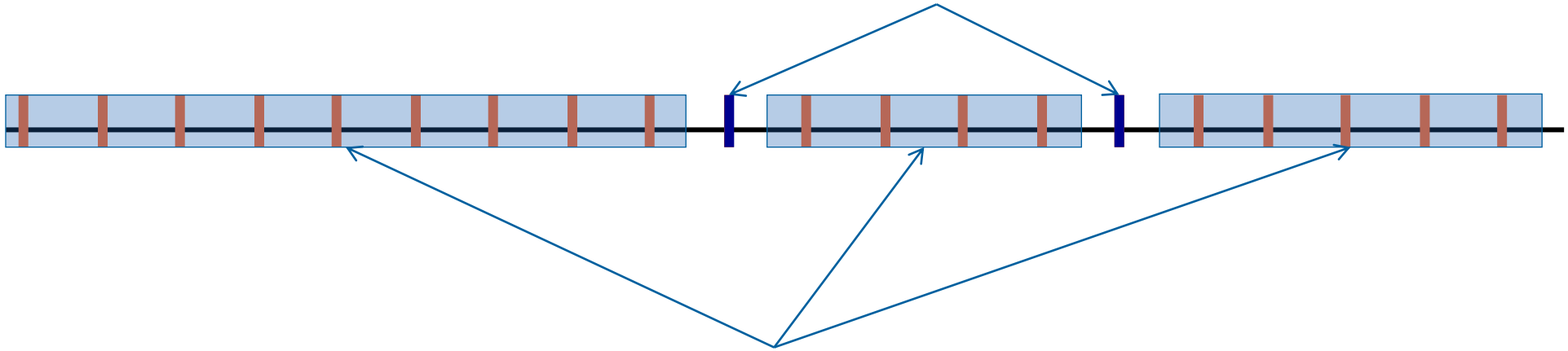
LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, 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 & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

Lonely SNPs [no LD]



LD blocks

Lonely SNPs [no LD]

LD blocks

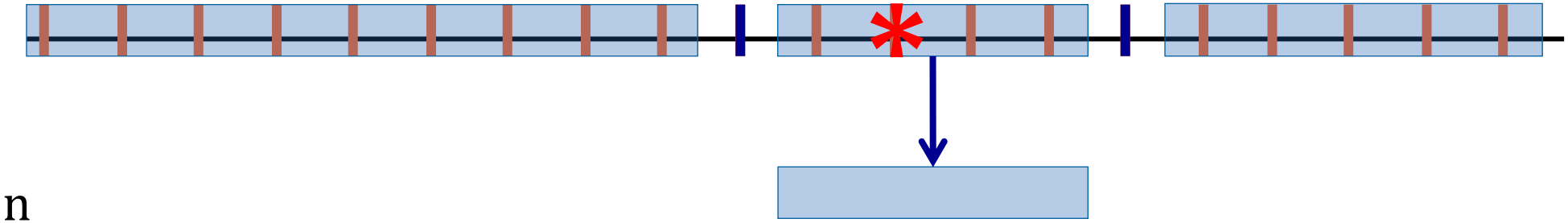
* Causal variants

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, 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 & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015



Association

All markers correlated with a causal variant show association

Lonely SNPs [no LD]

LD blocks

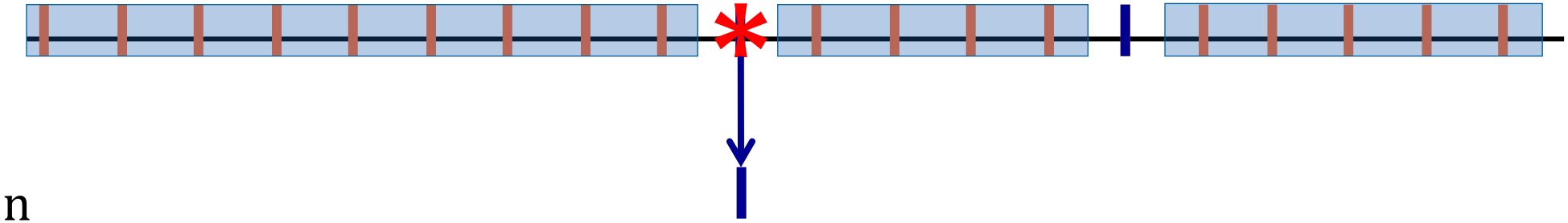
* Causal variants

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, 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 & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015



Association

Lonely SNPs only show association if they are causal

Lonely SNPs [no LD]

LD blocks

* Causal variants

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, 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 & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

Assuming a uniform prior, we see SNPs with more LD friends showing more association

The more you tag, the more likely you are to tag a causal variant

Lonely SNPs [no LD]

LD blocks

* Causal variants

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, 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 & Benjamin M Neale

Affiliations | Contributions | Corresponding author

Nature Genetics **47**, 291–295 (2015) | doi:10.1038/ng.3211
Received 07 March 2014 | Accepted 07 January 2015 | Published online 02 February 2015

Under pure drift we expect LD to have no relationship to differences in allele frequencies between populations

Estimating SNP heritability and Controlling Type 1 Error

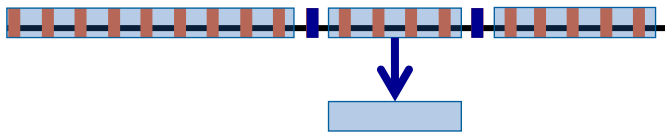
Confounders

SNP heritability

$$E[\chi_j^2] = 1 + \boxed{Na} + \frac{\boxed{h_g^2} N}{M} l_j$$

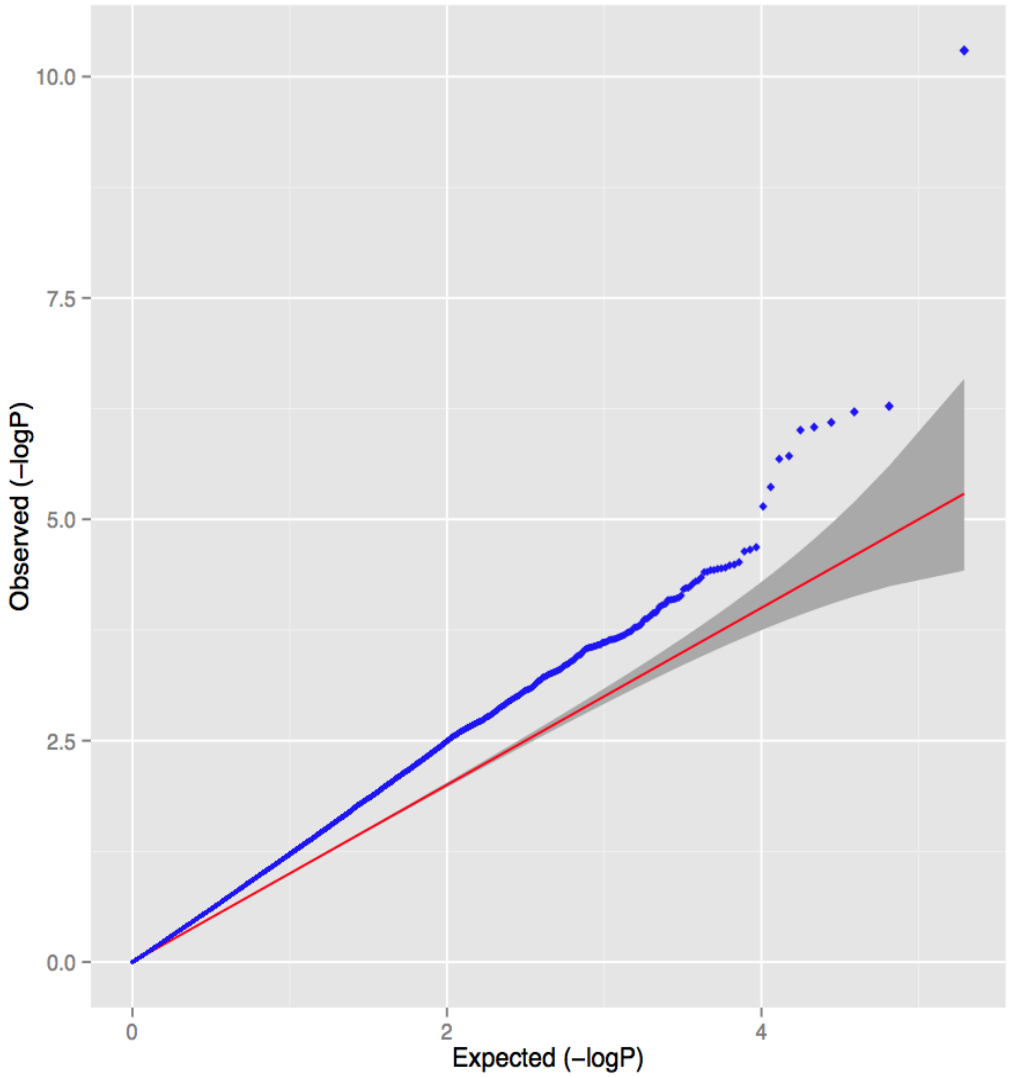
where N=sample size, M=# of SNPs, a=inflation due to confounding, h^2g is heritability (total obs.) and l_j is the *LD Score*

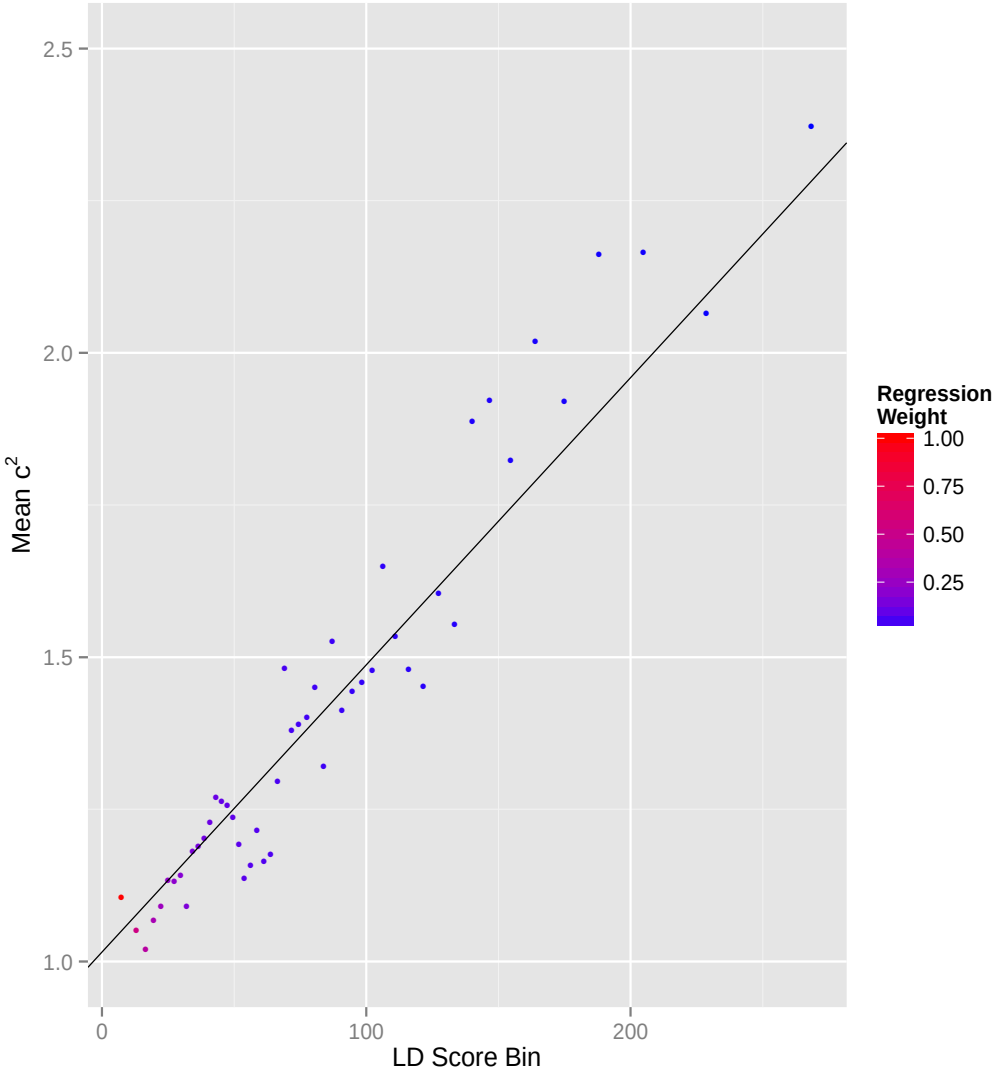
$$l_j = \sum_{k \neq j} r_{jk}^2$$



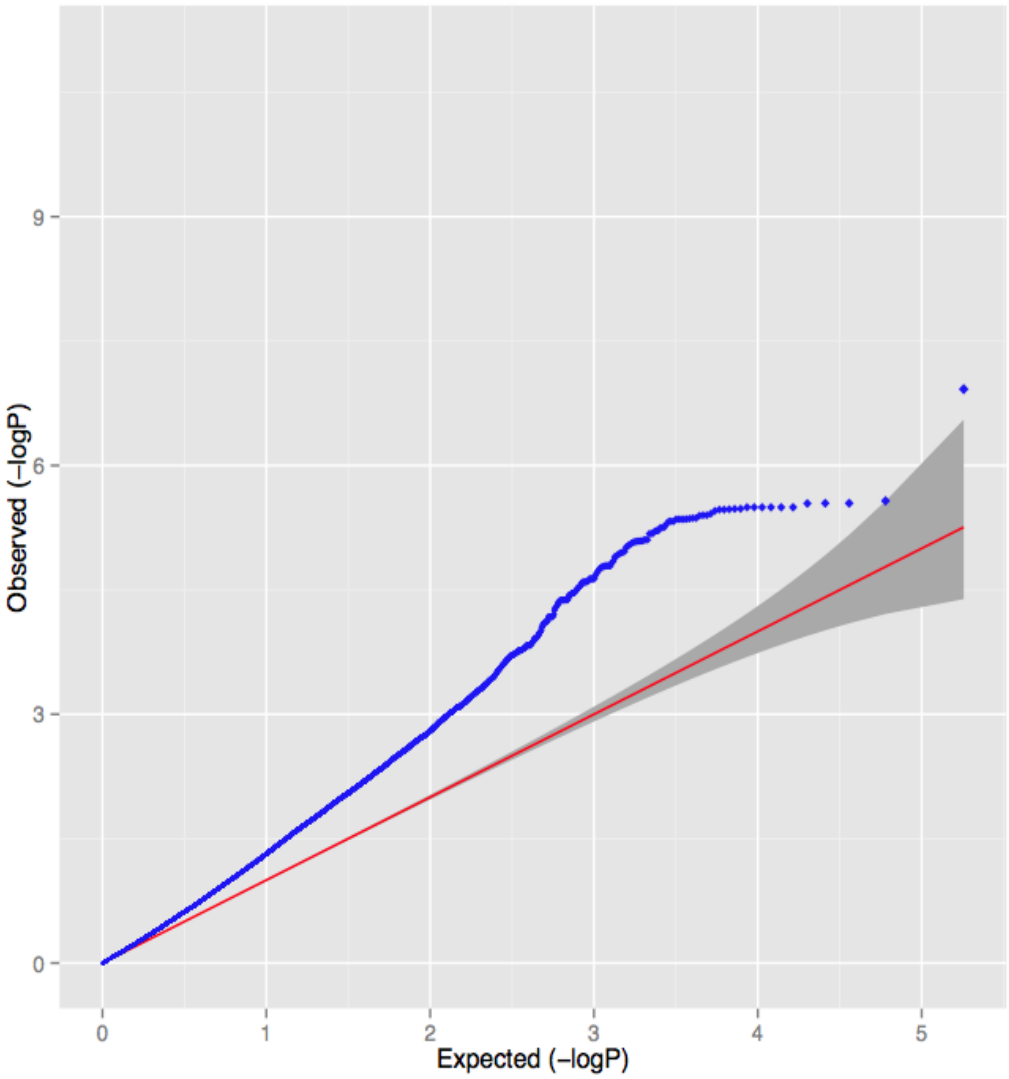
Bulik-Sullivan et al. Nature Genetics 2015
Yang et al. EJHG 2011

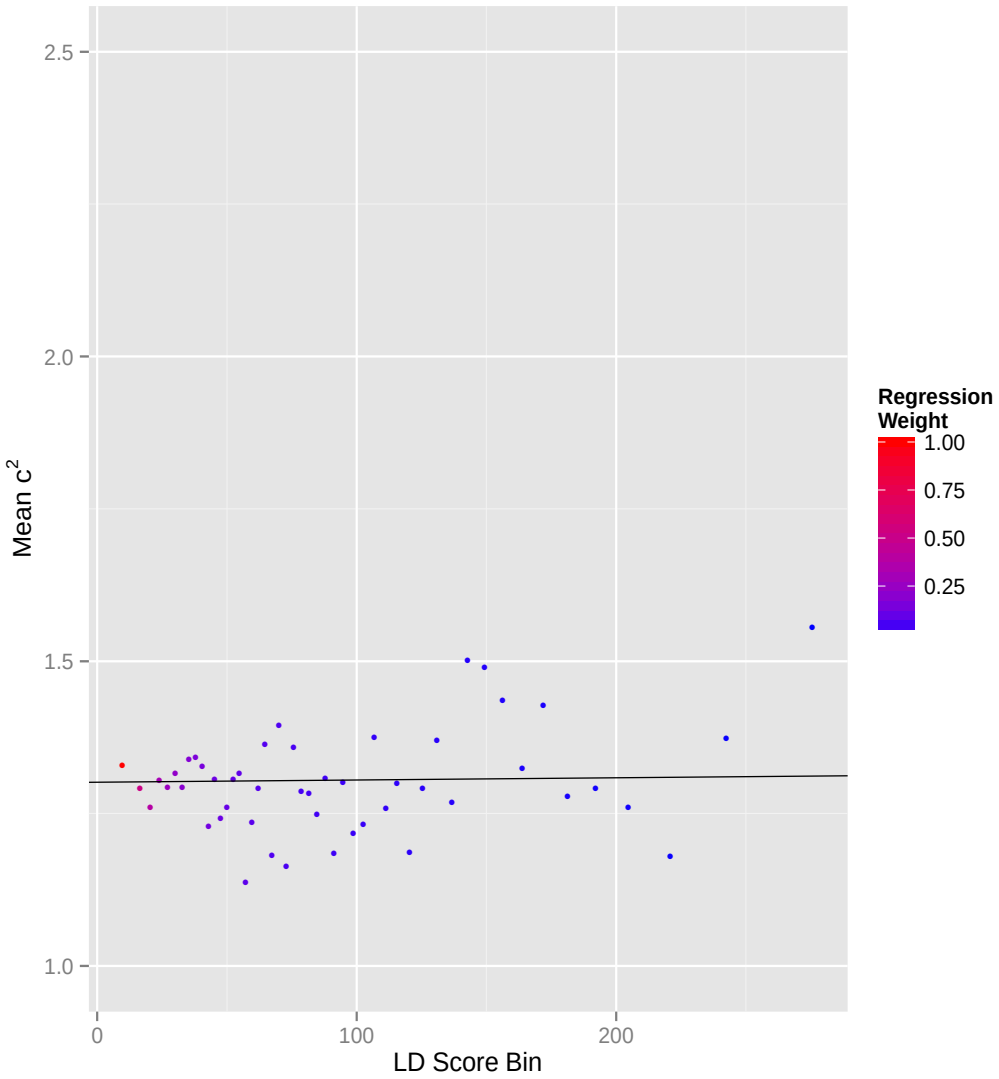
Simulated Polygenic Architecture





English Controls vs Swedish Controls





Estimating Genetic Covariance by Bivariate LD Score Regression

More precisely, under a polygenic model[11,13], the expected value of $z_{1j}z_{2j}$ for a SNP j is

Genetic covariance

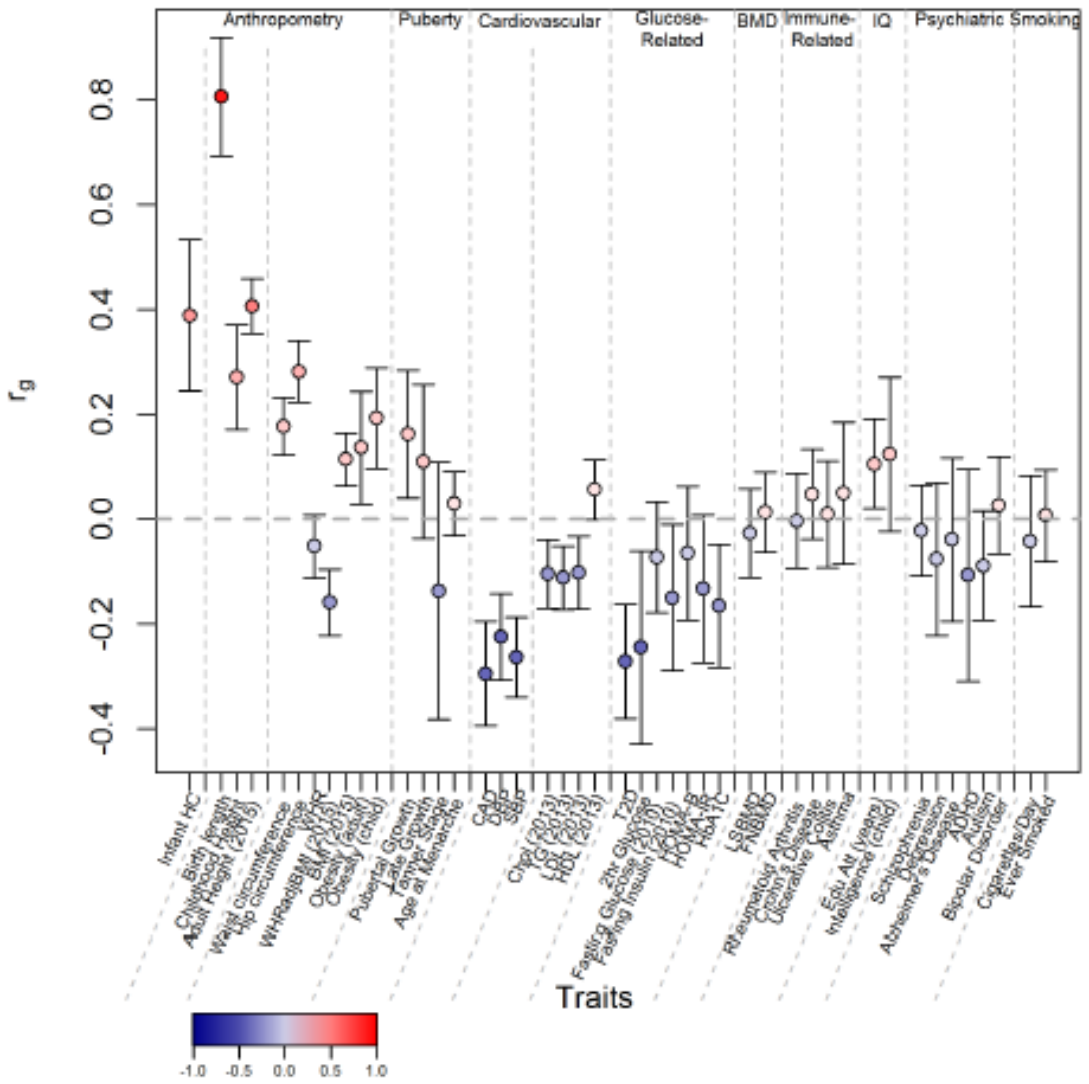
$$E[z_{1j}z_{2}\ell_j] = \frac{\sqrt{N_1N_2}\,\varrho_g}{M}\ell_j + \frac{\varrho N_s}{\sqrt{N_1N_2}} \qquad (1)$$

Sample overlap etc

where N_i is the sample size for study i, ϱ_g is the genetic covariance (defined in the Online Methods), ℓ_j is the LD Score[19], N_S is the number of individuals included in both studies and ϱ is the phenotypic correlation among the N_S overlapping samples. We

Bulik-Sullivan et al. *Nature Genetics* (2015)

Pervasive Genetic Pleiotropy



Horikoshi et al. *Nature* (2016)

Genomic SEM

Genomic SEM

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0566-x

Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits

Andrew D. Grotzinger[1*], Mijke Rhemtulla[2], Ronald de Vlaming[3,4], Stuart J. Ritchie[5,6], Travis T. Mallard[1], W. David Hill[5,6], Hill F. Ip[7], Riccardo E. Marioni[5,8], Andrew M. McIntosh[5,9], Ian J. Deary[5,6], Philipp D. Koellinger[3,4], K. Paige Harden[1,10], Michel G. Nivard[7,11] and Elliot M. Tucker-Drob[1,10,11]

Grotzinger

Nivard

Tucker-Drob

GenomicSEM

- Apply structural equation model to estimated genetic covariance matrices
  - Moves past family-based methods by allowing user to examine traits that could not be measured in the same sample

- Genomic SEM provides flexible framework for estimating limitless number of structural equation models using multivariate genetic data from GWAS summary statistics
  - Can be applied to sum stats with varying and unknown degrees of overlap

Genomic SEM uses these principles to fit structural equation models to genetic covariance matrices derived from GWAS summary statistics using 2 Stage Estimation

- Stage 1: Estimate Genetic Covariance Matrix and associated matrix of standard errors and their co-dependencies
  - We use LD Score Regression, but any method for estimating this matrix (e.g. GREML) and its sampling distribution can be used

- Stage 2: Fit a Structural Equation Model to the Matrices from Stage 1

Start with GWAS Summary Statistics for the Phenotypes of Interest

- No need for raw data
- No need to conduct a primary GWAS yourself: Download them online!
  - sumstats for over 3700 phenotypes have been helpfully indexed at http://atlas.ctglab.nl/
  - sumstats for over 4000 UK Biobank phenotypes are downloadable at http://www.nealelab.is/uk-biobank

CHR	SNP	BP	A1	A2	INFO	OR	SE	P	Nca	Nco	MAF
8	rs62513865	101592213	T	C	0.957	1.01461	0.0153	0.3438	59851	113154	0.07330
8	rs79643588	106973048	A	G	0.999	1.02122	0.0136	0.1231	59851	113154	0.09200
8	rs17396518	108690829	T	G	0.980	1.00331	0.0080	0.6821	59851	113154	0.43500
8	rs6994300	102569817	A	G	0.466	0.88126	0.4243	0.7658	16823	25632	0.00556
8	rs138449472	108580746	A	G	0.734	0.97181	0.0598	0.6320	41253	79756	0.00852
8	rs983166	108681675	A	C	0.991	0.99144	0.0080	0.2784	59851	113154	0.43200

Stage 1 Estimation: Multivariable LDSC

Create a genetic covariance matrix, S: an "atlas of genetic correlations"

Diagonal elements are (heritabilities)

$$S = \begin{bmatrix} h_1^2 & & & \\ \sigma_{g1,g2} & h_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{g1,gk} & \sigma_{g2,gk} & \cdots & h_k^2 \end{bmatrix}$$

Off-diagonal elements are coheritabilities

Stage 1 Estimation: Multivariable LDSC

Also produced is a second matrix, V, of squared standard errors and the dependencies between estimation errors

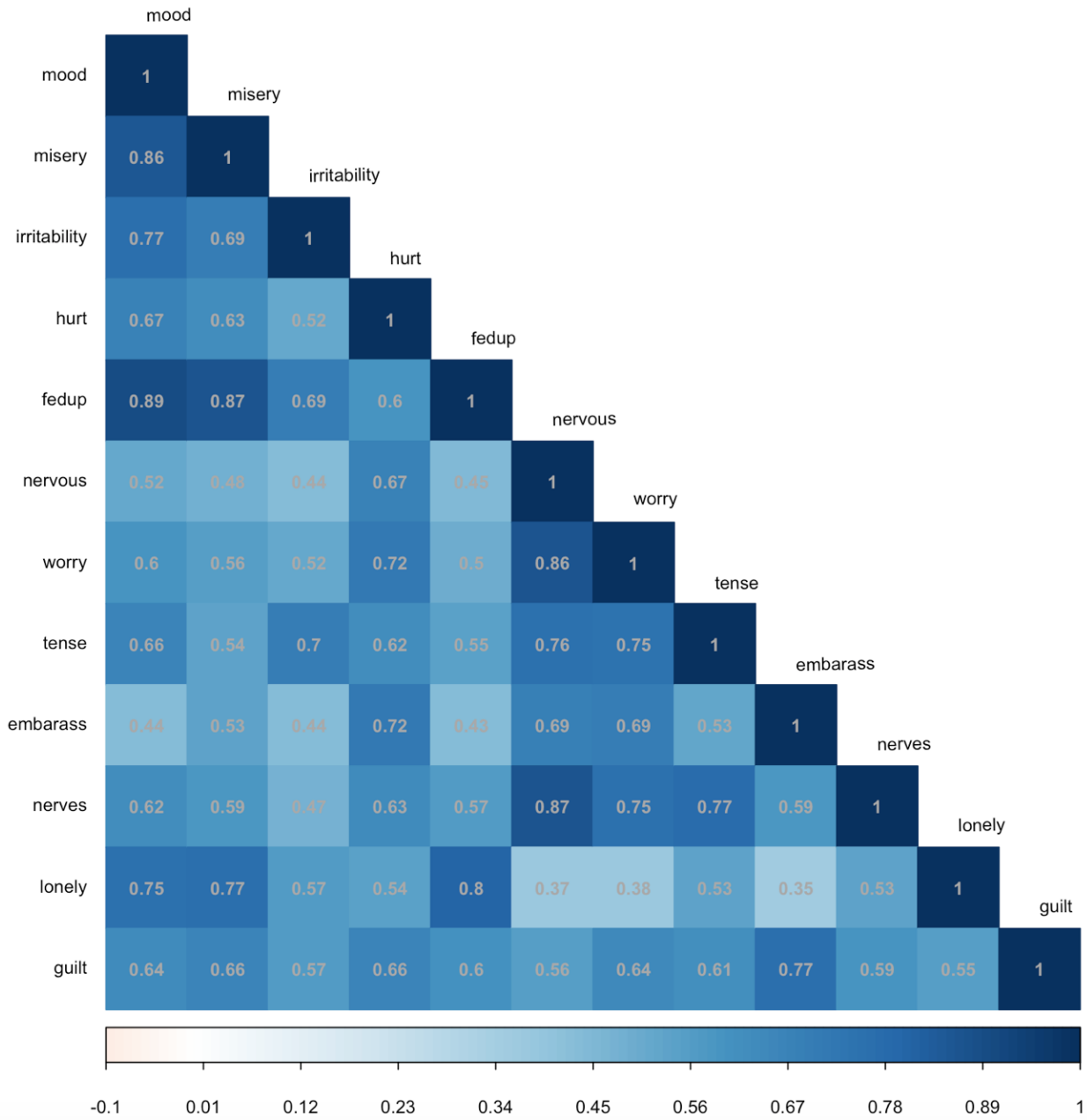
$$V_S = \begin{bmatrix} SE(h_1^2)^2 & & & & & & & \\ \mathrm{cov}(h_1^2, \sigma_{g1,g2}) & SE(\sigma_{g1,g2})^2 & & & & & & \\ \vdots & \vdots & \ddots & & & & & \\ & & & SE(\sigma_{g1,gk})^2 & & & & \\ & & & & \ddots & & & \\ & & & & & SE(h_j^2)^2 & & \\ & & & & & & \ddots & \\ & & & & & & & SE(\sigma_{gj,gk})^2 \\ & & & & & & & & SE(h_k^2)^2 \end{bmatrix}$$

Diagonal elements are squared standard errors of genetic variances and covariances

Off-diagonal elements are dependencies between estimation errors used to directly model dependencies that occur due to sample overlap from contributing GWASs

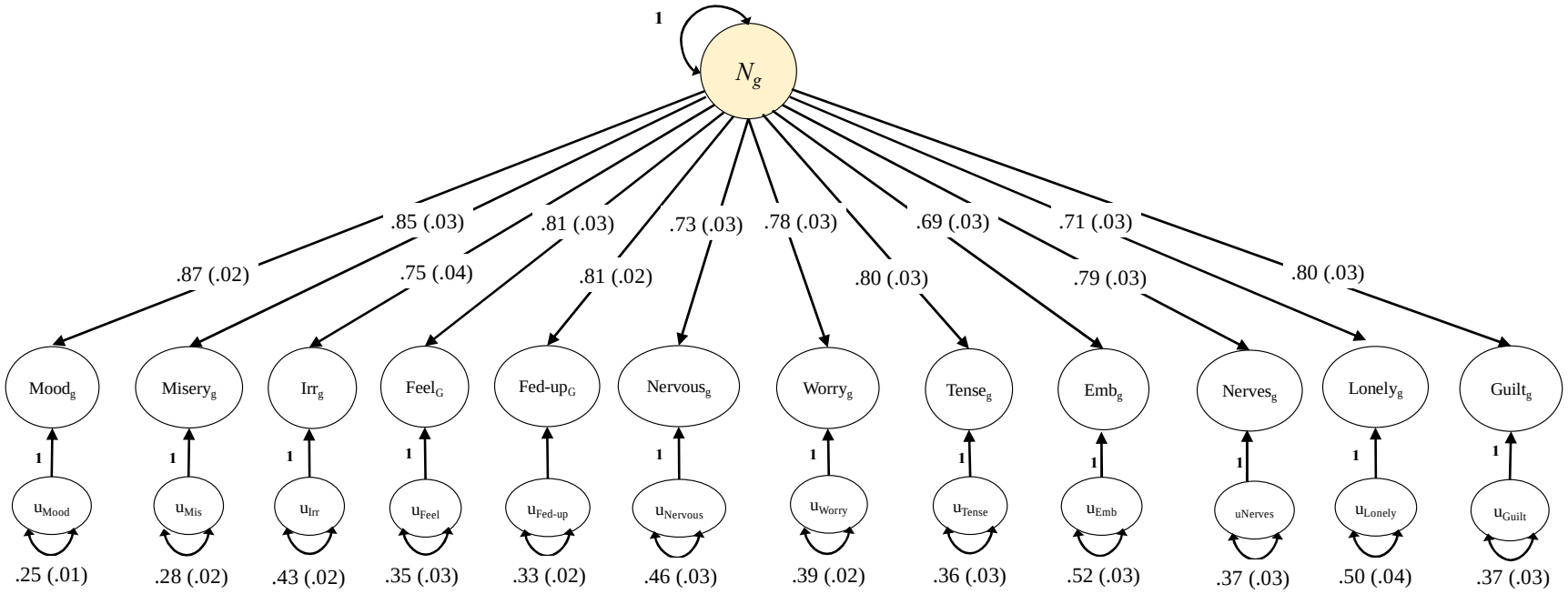
Genomic SEM Applications

Genetic Correlation Matrix



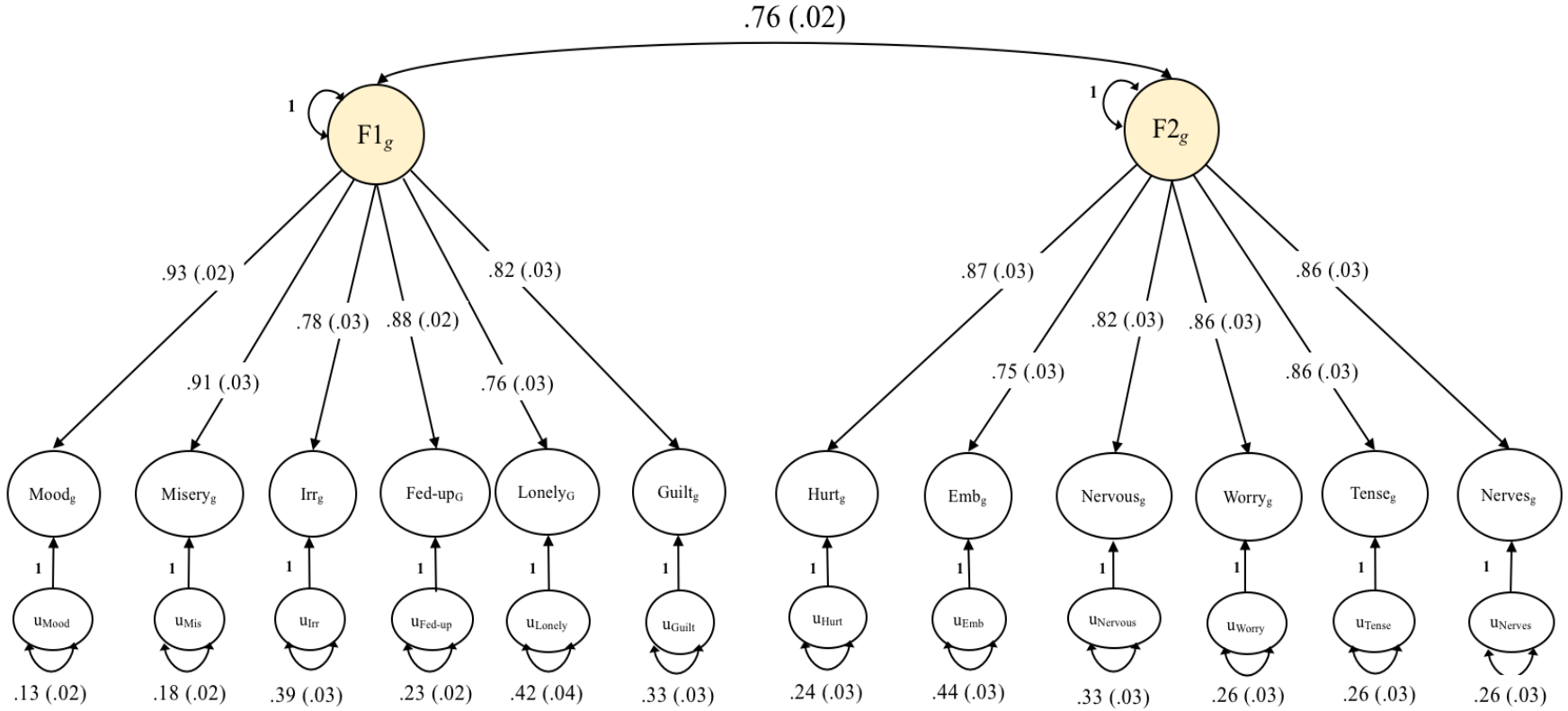
Model 1: Common Factor Model

chisq	df	p_chisq	AIC	CFI	SRMR
4884.104	54	0	4932.104	0.8933184	0.1095286



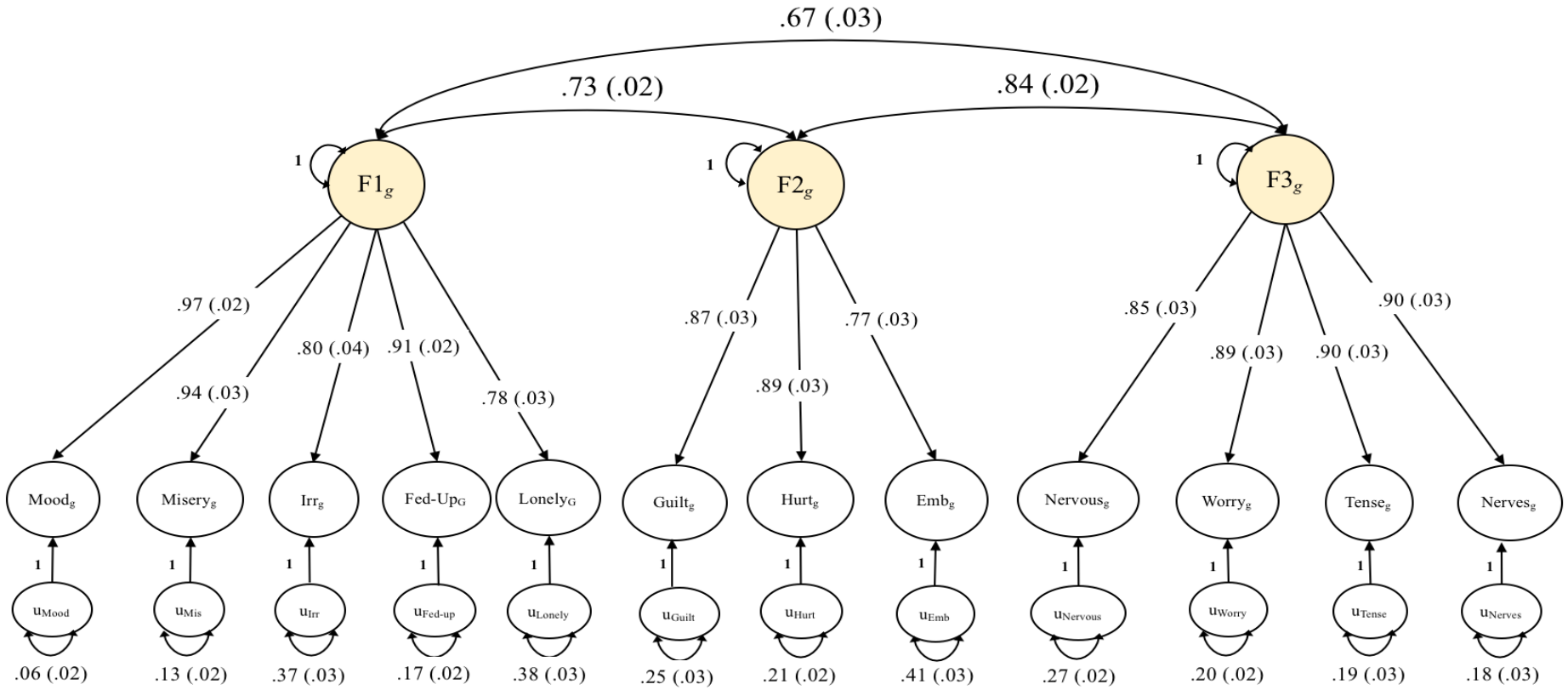
Model 2

chisq	df	p_chisq	AIC	CFI	SRMR
2758.176	53	0	2808.176	0.9402513	0.0766612

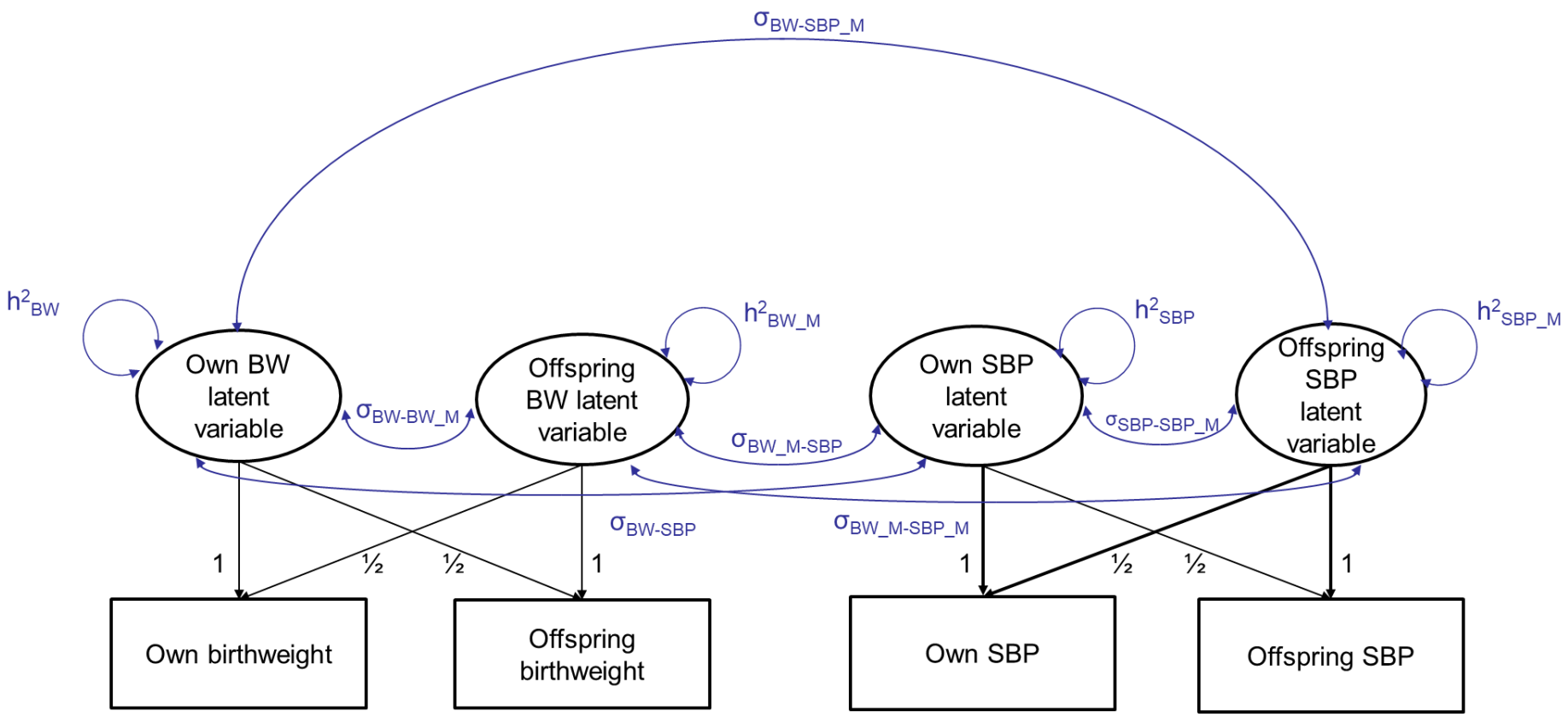


Model 3

chisq	df	p_chisq	AIC	CFI	SRMR
1879.308	51	0	1933.308	0.9596185	0.05733665



Example: Partitioning into Maternal and Fetal Components



Moen et al. *Behav Genet* (submitted)

Adding SNPs to Genomic SEM

Example: the *p* factor as a GWAS target

The American Journal of
Psychiatry
175th Year of Publication

REVIEWS AND OVERVIEWS
Mechanisms of Psychiatric Illness

All for One and One for All: Mental Disorders in One Dimension

Avshalom Caspi, Ph.D., Terrie E. Moffitt, Ph.D.

The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders?

Clinical Psychological Science
2014, Vol. 2(2) 119–137

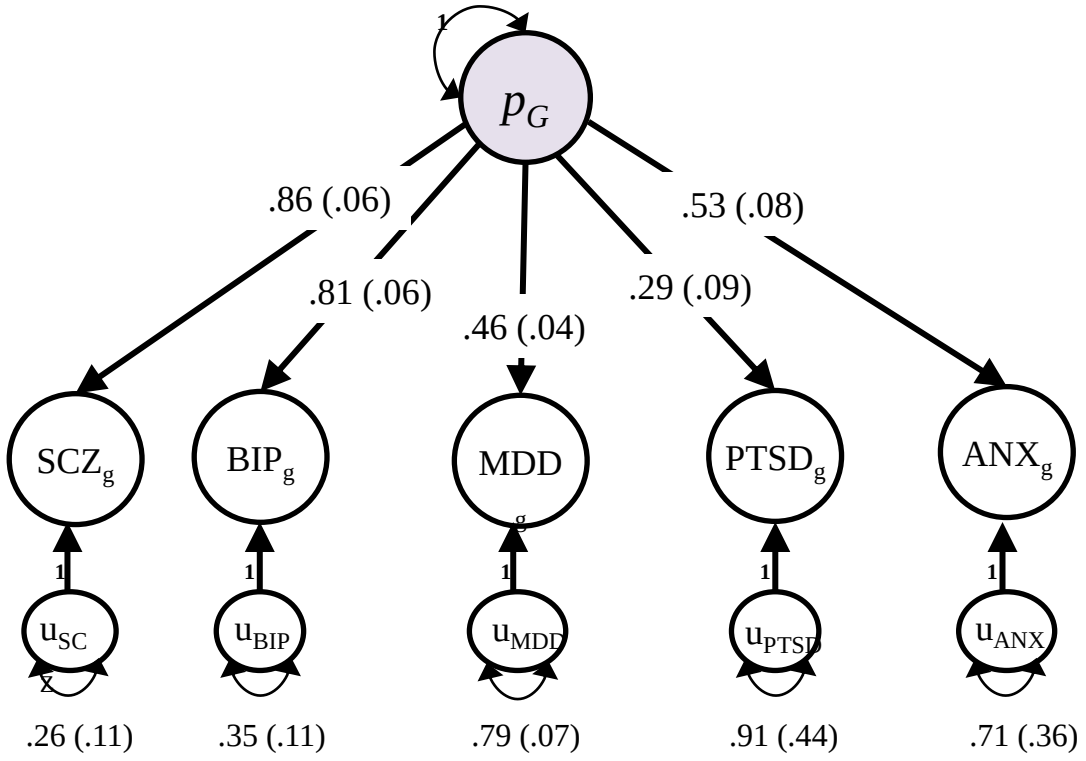
DOI: 10.1177/2167702613497473
cpx.sagepub.com
SAGE

Common Factor Model

Genetic Correlation Matrix

	SCZ	BIP	MDD	PTSD	ANX
SCZ	1				
BIP	0.74	1			
MDD	0.38	0.29	1		
PTSD	0.18	0.03	0.55	1	
ANX	0.38	0.23	0.91	0.09	1

0 0.11 0.22 0.33 0.44 0.55 0.66 0.77 0.88 0.99 1.1



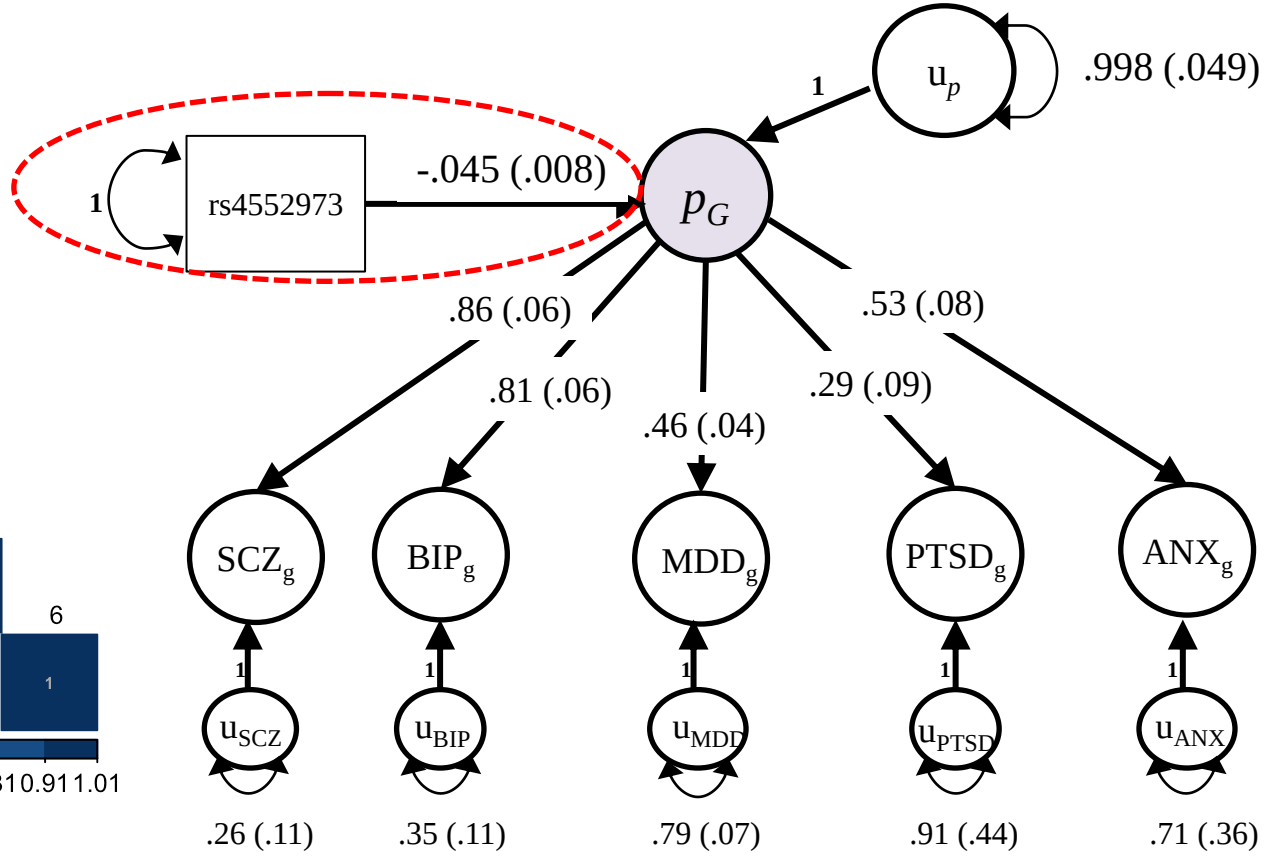
Add SNP Effects to the "Atlas"

$$S_{\text{Full}} = \begin{bmatrix} \sigma^2_{\text{SNP}} & & & & & \\ \sigma_{\text{SNP,g1}} & h_1^2 & & & & \\ \sigma_{\text{SNP,g2}} & \sigma_{\text{g1,g2}} & h_2^2 & & & \\ \sigma_{\text{SNP,g3}} & \sigma_{\text{g1,g3}} & \sigma_{\text{g2,g3}} & h_3^2 & & \\ \vdots & \vdots & & & \ddots & \\ \sigma_{\text{SNP,g}k} & \sigma_{\text{g1,g}k} & \sigma_{\text{g2,g}k} & \sigma_{\text{g3,g}k} & \cdots & h_k^2 \end{bmatrix}$$

Genetic Covariances from LDSC

Betas from GWAS sumstats scaled to covariances using MAFs

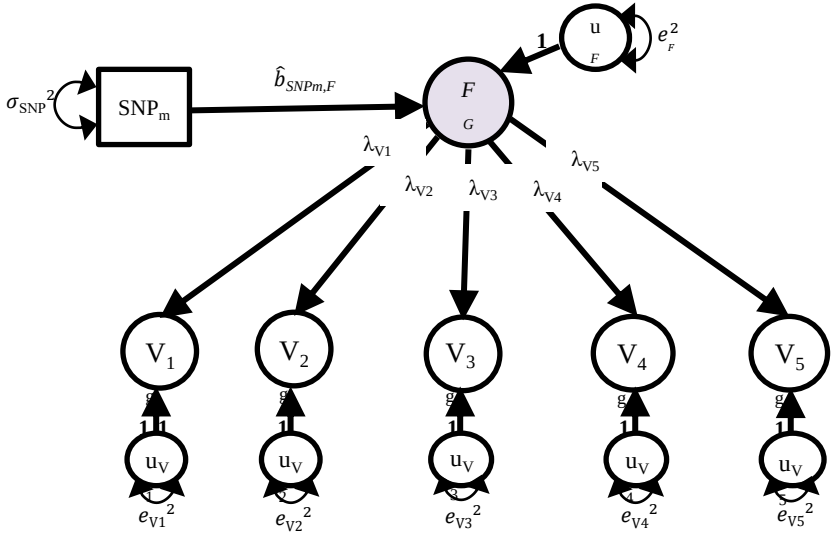
GWAS of a Latent Factor

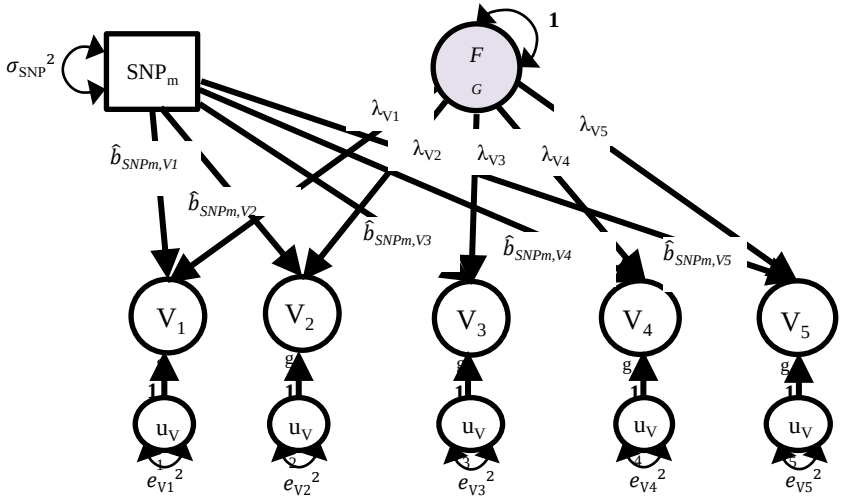


Estimates of SNP level heterogeneity (Q_{SNP})

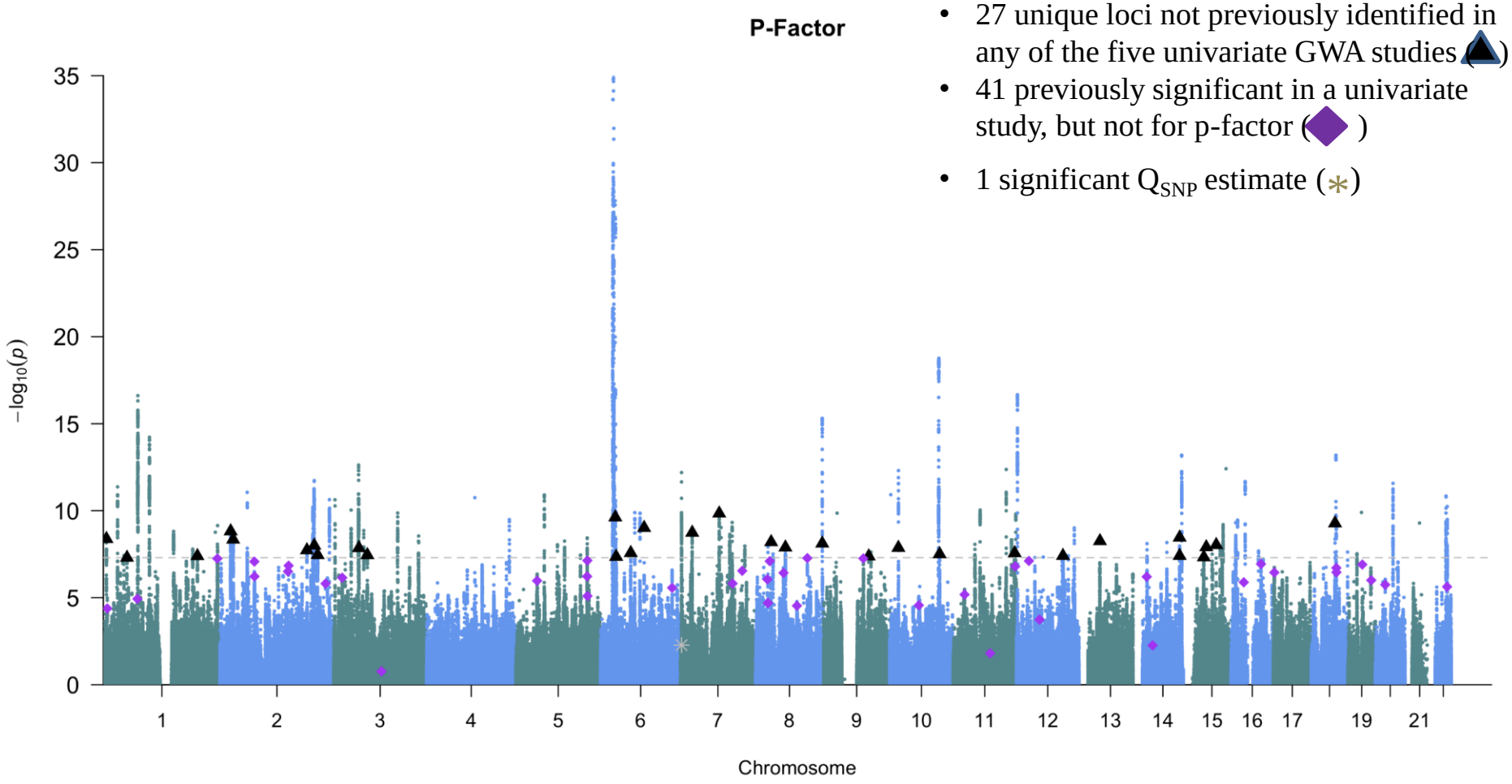
- Asks to what extent the effect of the SNP operates through the common factor

- χ^2 distributed test statistic, indexing fit of the common pathways model against independent pathways model



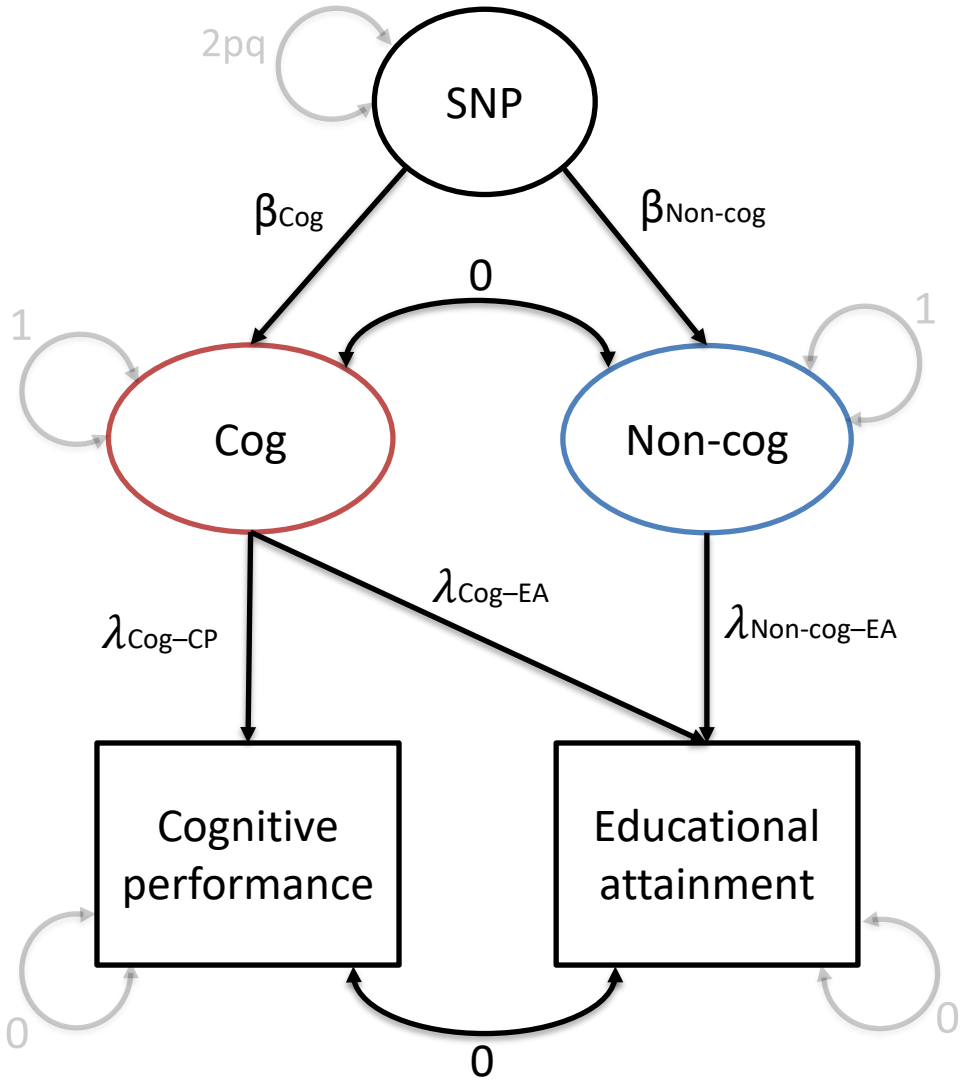


Manhattan Plot
Latent Factor

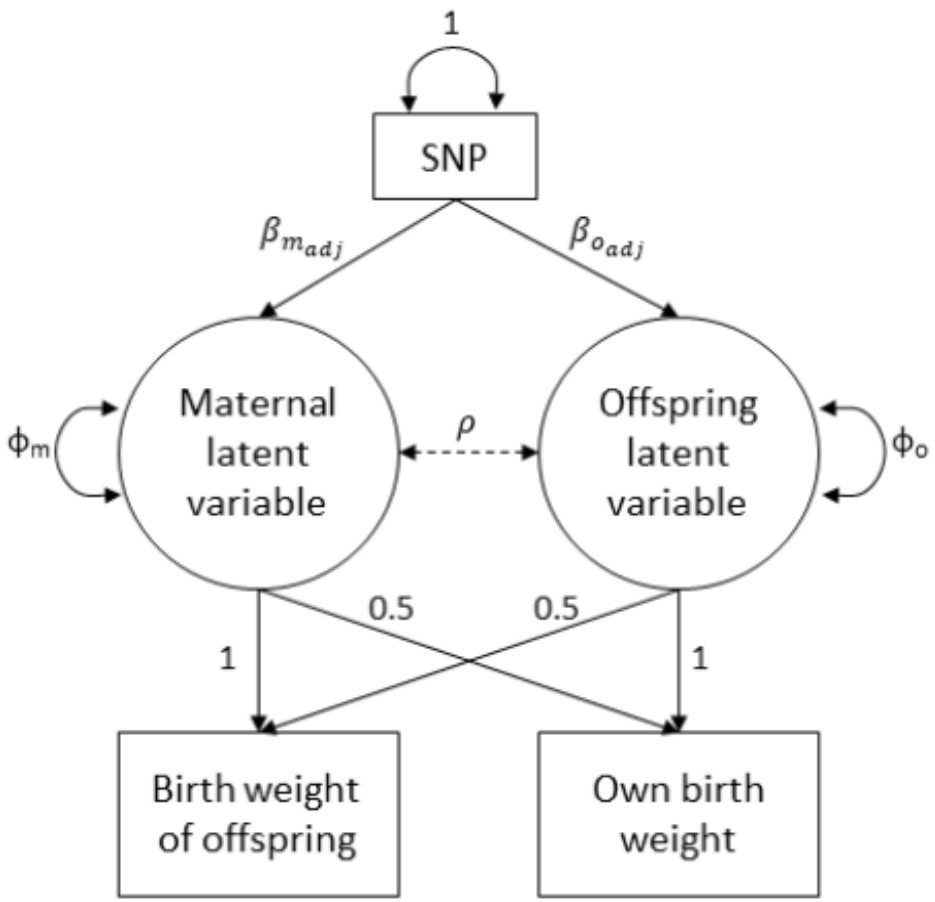


- 128 lead SNPs
- 27 unique loci not previously identified in any of the five univariate GWA studies (▲)
- 41 previously significant in a univariate study, but not for p-factor (◆)
- 1 significant Q_{SNP} estimate (✳)

GWAS by subtraction



Partitioning Genetic Effects



Warrington et al. (2021) *Nat Commun*

Further Reading

- Bulik-Sullivan B. et al (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, 47(3), 291-295.
- Bulik-Sullivan B. et al (2015). An atlas of genetic correlations across human diseases and traits. *Nat Genet*, 11, 1236-41.
- Demange PA. et al (2021). Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction. *Nat Genet*, 53(1), 35-44.
- Grotzinger A. et al (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav*, 3(5), 513-525.
- Warrington NM. et al (2021). Estimating direct and indirect genetic effects on offspring phenotypes using genome-wide summary results data. *Nat Commun*, 12(1), 5420.