Bivariate Gaussian Mixture Model of GWAS (BGMG) quantifies polygenic overlap between complex traits beyond genetic correlation

Oleksandr Frei[1*], Dominic Holland[2,3&], Olav B. Smeland[1,4&], Alexey A. Shadrin[1], Chun Chieh Fan[2,5,6], Aree Witoelar[1], Yunpeng Wang[1,2,6], Srdjan Djurovic[7,8], Wesley Thompson[9,10], Ole A. Andreassen[1,4], Anders M. Dale[2,3,6,11*]

Author affiliations

[1]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, 0424 Oslo, Norway

[2]Center for Multimodal Imaging and Genetics, University of California at San Diego, La Jolla, CA 92037, USA

[3]Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA

[4]Division of Mental Health and Addiction, Oslo University Hospital, 0407 Oslo, Norway

[5]Department of Cognitive Sciences, University of California at San Diego, La Jolla, CA 92093, USA

[6]Department of Radiology, University of California, San Diego, La Jolla, CA 92093, USA

[7]Department of Medical Genetics, Oslo University Hospital, 0424, Oslo, Norway

[8]NORMENT, KG Jebsen Centre for Psychosis Research, Department of Clinical Science, University of Bergen, 5020 Bergen, Norway

[9]University of California, San Diego, Department of Family Medicine and Public Health

[10]Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Capital Region of Denmark

[11]Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA

[&]These authors contributed equally to this work.

[*] To whom correspondence ought to be addressed:

oleksandr.frei@medisin.uio.no, andersmdale@gmail.com

ABSTRACT

Accumulating evidence from genome wide association studies (GWAS) suggests abundant presence of shared genetic influences among complex human traits and disorders. A major challenge that often limits our ability to detect and quantify shared genetic variation is that current methods of cross-trait analysis are not designed to work in scenarios with low or absent genetic correlation. Here we introduce a statistical tool BGMG (Bivariate Gaussian Mixture Model of GWAS) which can uncover various scenarios of genetic overlap regardless of genetic correlation, using GWAS summary statistics from studies with potentially shared participants. We perform extensive simulation on synthetic GWAS data to ensure that BGMG provides accurate estimates of model parameters in the presence of realistic linkage disequilibrium (LD) structure.

INTRODUCTION

In recent years, genome-wide association studies (GWASs) have successfully detected genetic variants associated with multiple complex human traits or disorders, providing important insights into human biology[1]. Understanding the degree to which complex human phenotypes share genetic influences is critical for identifying the etiology of phenotypic relationships, which can inform disease nosology, diagnostic practice and improve drug development. Most complex human phenotypes are known to have a polygenic architecture, i.e. their variation is influenced by many genetic variants. Given the large number of human phenotypes and the finite number of causal genetic variants, many variants are expected to influence more than one phenotype (i.e. exhibit allelic pleiotropy)[2,3]. This has led to cross-trait analyses quantifying genetic overlap becoming a widespread endeavor in genetic research, made possible by the public availability of most GWAS summary statistics (p-values and z-scores)[4,5].

Currently, the prevailing measure to quantify genetic overlap is genetic correlation. The square of genetic correlation gives the proportion of variance that the two traits share due to genetic causes. The sign of genetic correlation indicates whether genetic effects in both traits are, predominantly, sharing the same or the opposite effect direction. Genetic correlation can be quantified from raw genotypes using restricted maximum likelihood[6] or polygenic risk scores[7,8]; from a set of single-nucleotide polymorphisms (SNPs) that pass genome-wide significance threshold using Mendelian Randomization[9]; or from all SNPs, including those that do not reach genome-wide significance using Cross-Trait Linkage Disequilibrium Score Regression, LDSR[10]. A limitation to all these methods, however, is their inability to capture mixtures of effect directions across shared genetic variants. They only report overall positive, negative or no genetic correlation. Recent analyses suggest that across traits the correlation in the directionality and size of SNP effects is not the same for all mutually associated SNPs[10] (see Fig 1 for different cross-trait genetic relationships). This is exemplified by the genetic relationship between schizophrenia and cognitive function[11]. Despite consistent estimates of a negative genetic correlation between schizophrenia and different cognitive traits[12,13], a

minority of genetic loci associated with both schizophrenia and cognitive traits show schizophrenia risk alleles associated with *higher* cognitive performance[11]. Moreover, recent analyses suggest that different complex phenotypes are influenced by different numbers of causal variants, i.e., some traits are more polygenic than others[14], a concept in line with the endophenotype hypothesis[15]. To improve our understanding of the polygenic architecture of complex traits and their intricate relationships, new statistical tools are needed for quantification of genetic overlap.

Here we introduce the Bivariate Gaussian Mixture Model for GWAS (BGMG), which provides a measure of genetic overlap expressed as the proportion of SNPs associated with two traits. BGMG incorporates a causal mixture model[14] to yield a prior distribution of genetic effect sizes, and allows for overlapping samples. To estimate polygenic overlap, BGMG models true per-SNP effect sizes as a mixture of four bivariate normal distributions, illustrated in Fig 1: two causal components specific to each trait; one causal component of SNPs affecting both traits; and a null component of SNPs with no effect on either trait. Our statistical model provides a probability distribution function relating observed signed test statistics (GWAS z-scores) to the underlying per-SNP effect sizes, incorporating effects of LD structure, minor allele frequency, sample size, cryptic relationships, and sample overlap, to capture all these effects on GWAS z-scores. The parameters of the mixture model are estimated from the summary statistics by direct optimization of the likelihood function.

We show in simulations that our model differentiates cross-trait scenarios with no polygenic overlap from scenarios with significant polygenic overlap, regardless of genetic correlation. We also show that BGMG estimates of genetic correlation are consistent with estimates of cross-trait LDSR and are not affected by sample overlap. Altogether, the results demonstrate the feasibility of BGMG for quantifying shared polygenic components influencing complex human phenotypes.

RESULTS

Overview of BGMG model

BGMG is based on the idea that at the causal level only a certain proportion of SNPs is associated with a trait of interest, while the remaining SNPs have no effect (mixture model). In a joint analysis of two traits we expect some SNPs to have an effect on both traits; some SNPs to have an effect on one trait but not the other; and the majority of SNPs to have no effect on either trait. In this context, the term "polygenic overlap" signifies the fraction of causal variants shared between traits, exceeding that expected by chance given the polygenicity of those traits. Based on these assumptions, BGMG models additive genetic effect sizes β_{1j}, β_{2j} of SNP j on the two traits as a mixture of four bivariate Gaussian components:

$$(\beta_{1j}, \beta_{2j}) \sim \pi_0 N(0,0) + \pi_1 N(0, \Sigma_1) + \pi_2 N(0, \Sigma_2) + \pi_{12} N(0, \Sigma_{12}), \qquad (1)$$

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \ \Sigma_{12} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad (2)$$

where π_1 and π_2 are weights of the two components affecting the first and the second trait; π_{12} is a weighting of the component affecting both traits; and π_0 is the fraction of SNPs that are null (non-causal) for both traits, $\pi_0 + \pi_1 + \pi_2 + \pi_{12} = 1$; σ_1^2 and σ_2^2 control expected magnitudes of per-variant effect sizes; and ρ_{12} is the coefficient of genetic correlation, which is calculated from the subset of SNPs affecting both traits (see the Online Methods). Genome-wide genetic correlation r_g is related to the parameter ρ_{12} as $r_g = \rho_{12}\pi_{12}/\sqrt{\pi_1^u \pi_2^u}$, where π_1^u and π_2^u indicate total (univariate) proportion of causal SNPs in each of the two traits ($\pi_1^u = \pi_{12} + \pi_1$, $\pi_2^u = \pi_{12} + \pi_2$). All parameters are assumed to be the same for all SNPs.

From the prior probability density function for association coefficients (β_{1j}, β_{2j}), we derive the likelihood term for observed GWAS signed test statistics, incorporating: effects of linkage disequilibrium structure (allelic correlation r_{ij} between variants i and j); heterozygosity ($H_j = 2p_j(1-p_j)$ where p_j is the minor allele frequency of the j-th variant); number of subjects genotyped per variant (N_{1j} and N_{2j}); inflation due to cryptic relatedness σ_{01}^2 and σ_{02}^2; and inflation due to sample overlap ρ_0. Specifically (see Supplementary Note),

$$(z_{1j}, z_{2j}) = (\delta_{1j}, \delta_{2j}) + N\left((0,0), \begin{bmatrix} \sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02} \\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2 \end{bmatrix}\right), \qquad (3)$$

$$\delta_{\cdot j} = \sqrt{N_{\cdot j}} \sum_i \sqrt{H_i} r_{ij} \beta_{\cdot j}$$

The nine parameters of the model $(\pi_1, \pi_2, \pi_{12}, \sigma_1^2, \sigma_2^2, \rho_{12}, \sigma_{01}^2, \sigma_{02}^2, \rho_0)$ are fit by direct optimization of the weighted log likelihood, with weights inversely proportional to the LD score[16]. Confidence intervals for all parameters are estimated from the Observed Fishers Information matrix.

Forcing $\pi_{12} = 1$ (so that $\pi_0 = \pi_1 = \pi_2 = 0$) reduces our model to the same assumptions as in cross-trait LD score regression[10]. Under this constraint our model predicts that GWAS signed test statistics follow bivariate Gaussian distribution with zero mean and variance-covariance matrix

$$\Sigma_j = \ell_j \begin{bmatrix} N_{1j}\sigma_1^2 & \sqrt{N_{1j}N_{2j}}\rho_{12}\sigma_1\sigma_2 \\ \sqrt{N_{1j}N_{2j}}\rho_{12}\sigma_1\sigma_2 & N_{2j}\sigma_2^2 \end{bmatrix} + \begin{bmatrix} \sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02} \\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2 \end{bmatrix},$$

i.e., $(z_{1j}, z_{2j}) \sim N(0, \Sigma_j)$, where $\ell_j = \sum_i H_i r_{ij}^2$ is the LD score (adjusted for heterozygosity). This model is in perfect agreement with cross-trait LD score regression, with expected chi square statistics $E(z_{1j}^2)$, $E(z_{2j}^2)$ and cross-trait correlation $E(z_{1j}z_{2j})$ being proportional to the LD score of j-th SNP, and parameters ρ_0, σ_{01}, σ_{02} playing the role of LD score regression intercepts. The only distinction so far is that we choose to model effect sizes that are independent of allele frequency, leading to the incorporation of H_i in our model; this factor is absent from the LD score regression model due to the assumption there of effect sizes that are inversely proportional to H_i.

Relaxing the constraint $\pi_{12} = 1$, we find that estimating its value from the GWAS summary statistics never yields values exceeding $\hat{\pi}_{12} = 1\%$, across a wide range of complex traits. The fact that $\hat{\pi}_{12}$ is typically much less than 100% supports polygenic model and contradicts omnigenic model[17]. We show in simulations that under polygenic model the measure of π_{12}, i.e., the proportion of causal variants affecting both traits, has both sensitivity and specificity in measuring genetic overlap.

Using GWAS summary statistics we find that some traits have a strong component of shared genetic effects but no genetic correlation. I.e., π_{12} significantly different from zero, but ρ_{12} is not. A challenge in detection and interpretation of such cases is that for polygenic traits we always expect a substantial portion of SNPs to have elevated test statistics in both traits due to LD with causal variants, even if no SNPs are causally affecting either trait. In our model, we account for this by modeling genetic effects at the causal level, and by reporting π_{12} as significant if it exceeds polygenic overlap that we expect by chance from the product $\pi_1^u \pi_2^u$.

Simulations

We perform simulations to validate that the BGMG estimates of polygenic overlap (π_{12}) are not affected by spurious association signals arising in large LD blocks, and that BGMG estimates of genetic correlation are not inflated by sample overlap. The simulations also allow us to compare BGMG estimates of genetic correlation with those of LDSR.

In all simulations we obtain synthetic GWAS results for a panel of N = 100,000 samples ("individuals"), generated by HapGen2[18] using 1000 Genomes[19] data to approximate the LD structure for European ancestry. For each simulation run we generated two quantitative traits for each individual by drawing effect sizes (β_{1j}, β_{2j}) from the four component mixture model (1), varying polygenicity of each phenotype (π_1^u and π_2^u), and polygenic overlap (π_{12}). We choose polygenicity of each trait ranging from 10^{-3} (high polygenicity) to 10^{-4} (medium polygenicity) to 10^{-5} (low polygenicity). We also choose between perfect polygenic overlap ($\pi_{12} = \pi_1^u$), partial polygenic overlap set to 10% of polygenicity of the traits ($\pi_{12} = \pi_1^u/10$), and random polygenic overlap, which arise by chance if markers are spread randomly throughout the genome (independent prior probabilities, so that $\pi_{12} = \pi_1^u \pi_2^u$). In all simulations, we set narrow sense SNP heritability of each trait to $h^2 = 0.5$.

Fig. 2 illustrates BGMG components for the four synthetic scenarios, introduced in Fig 1, using $\pi_1^u = \pi_2^u = 0.01\%$ (medium polygenicity). The distribution of GWAS effect sizes, shown in Fig. 2, is different from distribution of the causal effect sizes shown in Fig. 1, and shows a much larger fraction of variants associated with the phenotype of interest, arising through LD.

Fig. 3 and Supplementary Table 1 show estimates of polygenicity and polygenic overlap on synthetic data with realistic LD structure, averaged across 10 simulation

runs. For high and medium polygenicity the resulting univariate estimates of polygenicity, $\hat{\pi}_1^u$ and $\hat{\pi}_2^u$, are biased downwards by approximately 10% (relative percentage w.r.t. the expected value). For low polygenicity the estimates are biased upwards by a factor of 3, consistently with previously published simulation results[14]. Despite the bias in polygenicity estimates, the relative size of polygenic overlap, $\hat{\pi}_{12}/\hat{\pi}_1^u$, is in a good agreement with expected values (1.0 for complete overlap, and 0.1 for partial overlap). For random overlap, BGMG overestimates $\hat{\pi}_{12}$ and the ratio $\hat{\pi}_{12}/\hat{\pi}_1^u$ has values around 0.01. Simulation results for traits with uneven polygenicity (10-fold difference, $\pi_1^u = 10\pi_2^u$, and 100-fold difference, $\pi_1^u = 100\pi_2^u$) show the same pattern as simulations with equal polygenicity ($\pi_1^u = \pi_2^u$), see Supplementary Table 2. All discrepancies in polygenicity estimates $\hat{\pi}_1^u$, $\hat{\pi}_2^u$ and $\hat{\pi}_{12}$ disappear in simulations without LD structure (see Supplementary Table 3).

In addition to point estimates of polygenicity we investigated BGMG performance using univariate quantile-quantile plots (QQ plots) on logarithmic scale, as shown in Supplementary Fig. 1 and 2. The advantage of QQ plots is that they emphasize behavior in the tails of a distribution, and provide a valuable visual aid in showing how well a model fits data. The overall QQ plot (Supplementary Fig 1) shows reasonably good fit across the entire range of p-values. QQ plots stratified by heterozygosity, H, and LD score, ℓ, (Supplementary Fig 2) show minor mismatches for low values of H and ℓ.

Fig. 4 compares BGMG and LDSR estimates of genetic correlation r_g and corresponding standard errors. These simulations cover scenarios with complete and partial polygenic overlap, using the same set of samples (N=100 000) to perform GWAS in both traits (i.e. we validate complete sample overlap). The results show that the estimates of genetic correlation from BGMG and LDSR are not biased by complete sample overlap. Error bars for BGMG are generally larger than LDSR because it estimates a larger set of parameters (9 parameters for BGMG model, 6 parameters for LDSR model).

DISCUSSION

Here we introduce BGMG, a new statistical tool for cross-trait analysis using GWAS summary statistics which builds on a Gaussian mixture modelling framework. BGMG offers two major advances compared to other currently available cross-trait analyses. First, BGMG allows for a mixture of same and opposite allelic effect directions among the shared genetic component, which is likely a pervasive occurrence given the large number of causal variants influencing traits and their distinct genetic etiologies[11-13,20,21]. Second, BGMG takes into account the unique polygenic architecture underlying each complex trait, which widely differs between traits, both in terms of the number of causal SNPs and the effect sizes of the causal SNPs[14,22,23]. Thus, BGMG enables a more complete quantification of polygenic overlap in various cross-trait scenarios than provided by other available tools estimating genetic overlap[7-10,24,25].

Using simulations, we show that BGMG robustly captures the degree of shared genetic components for various scenarios of polygenic overlap, with (Fig. 3) and without (Fig. 4) genetic correlation, and demonstrate that BGMG estimates are not inflated by sample overlap (Fig. 4). Simulation results also reveal a small bias in polygenicity estimates (Supplementary Tables 1-4), which we expect to fix in our future work by more elaborate handling of the LD structure and switching our simulation pipeline to the standard GWAS association model which includes genetic covariates.

The BGMG model is based on a causal mixture model which gives a biologically more plausible prior distribution of genetic effect sizes compared to the "infinitesimal" model applied by other available cross-trait analyses[23,26]. Causal mixture model is also known as a spike-and-slab distribution of effect sizes, and represents the most common way to perform simulations[10,16,27,28]. A notable strength of the spike-and-slab model is the significant improvement of polygenic risk scores[29-31].

Lack of correlation between two variables of course does not imply independence. By using simulated GWAS data, here we show that some scenarios of polygenic overlap not captured by genetic correlation tools are uncovered by BGMG (Fig. 4). BGMG controls for the probability that some SNPs will, by chance or due to LD structure given high polygenicity, be jointly associated with two traits. In addition to offering insights into shared genetic architectures of complex traits, the BGMG modelling framework can be used to improve power for SNP discovery by estimating the posterior effect size of SNPs associated with one trait given the test statistics in another trait. Moreover, we expect that more accurately estimated effect sizes will improve predictive power of polygenic risk scores.

The BGMG model has some limitations. First, the model assumes similar LD structure among studies and is not currently applicable for analysis across different ethnicities. Second, the model assumes that causal variants are uniformly distributed across the genome. We plan to address this limitation by incorporating genomic annotations, known to be differentially enriched for true associations[32]. Third, individual parameters of the mixture model might have lower estimation accuracy than their combinations – for example, we observe larger estimation errors for π_1 and σ_1^2 compared to the heritability estimate $h^2 \propto \pi_1 \sigma_1^2$ (due to inversely-correlated errors). Fourth, lack of significant estimates of polygenic overlap using BGMG does not exclude the possibility that some SNPs associations may indeed be linked to both traits, but less than expected by chance.

In conclusion, BGMG represents a useful addition to the tool-box for cross-trait GWAS analysis. By appropriately taking into account the intricate polygenic architectures of complex phenotypes BGMG allows for measures of polygenic overlap beyond genetic correlation. We expect this to lead to new insights into the pleiotropic nature of human genetic etiology. BGMG is available as a MATLAB/Octave package (see URLs; will be released after publication).

ACKNOWLEDGMENTS:

This work was supported by the Research Council of Norway (#223273, #225989, #248778) South-East Norway Health Authority (#2016-064, #2017-004), KG Jebsen Stiftelsen (#SKGJ-Med-008) and National Institutes of Health.

AUTHOR CONTRIBUTIONS:

Conceived and designed the study – AMD, OAA, OF, Method development – AMD, OF, DH, Analysis and interpretation of results – OF, DH, AAS, OBS, CCF, YW, AW, SD, Drafting manuscript – OF, OAA, OBS, Revision and approval of final manuscript – ALL AUTHORS

URLS:

https://github.com/precimed/ - BGMG code (MATLAB/Octave), will be released after publication

COMPETING FINANCIAL INTERESTS: The authors declare no competing financial interests.

CORRESPONDING AUTHORS:

Oleksandr Frei (oleksandr.frei@medisin.uio.no)
Anders M. Dale (andersmdale@gmail.com)

FIGURES AND TABLES

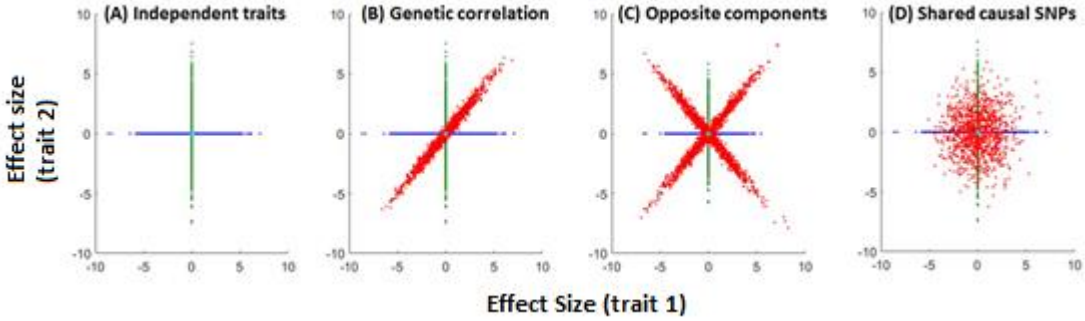

Figure 1: Components of causal mixture model in four extreme scenarios of polygenic overlap (synthetic data). Each point represents a SNP; horizontal and vertical axis show SNP causal effect sizes β_{1j}, β_{2j} on the first and on the second traits, respectively. Each component is simulated to have 0.01% of markers, randomly spread throughout the entire genome. Fig. **1A** shows a scenario where causal variants do not overlap between the two traits. Fig. **1B** shows an additional component of variants affecting both traits with the same (concordant) direction of effects. Fig. **1C** adds a fourth component of markers that affect the two phenotypes in the opposite (discordant) direction. Finally, Fig. **1D** shows a scenario similar to Fig. 1C, but with no clear separation between concordant and discordant components. In scenarios 1C and 1D genetic correlation is zero despite polygenic overlap.

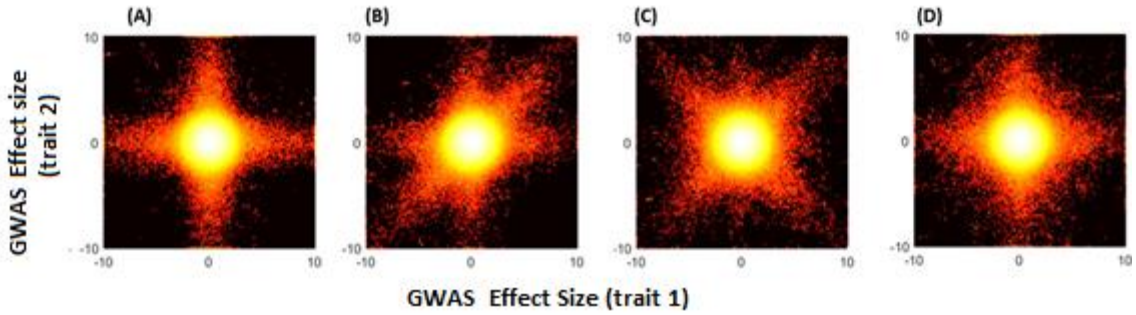

Figure 2: Density plot of the effect sizes from simulated GWAS data, N=100 000, using the same underlying causal mixture model as shown in Fig 1. Color indicates bivariate density of SNPs; horizontal and vertical axes show GWAS effect size estimate for the first and second traits, respectively.

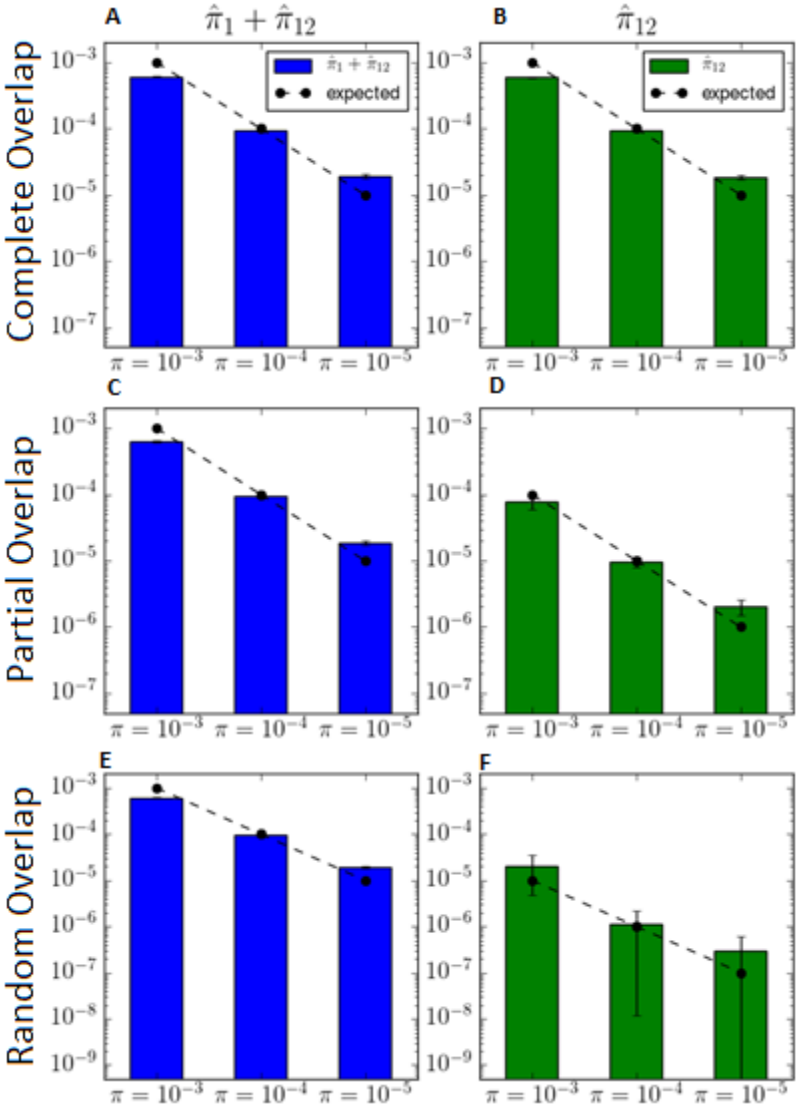

Figure 3: Estimates of polygenicity and polygenic overlap on synthetic data, in the absence of genetic correlation (ρ_{12} = 0). Rows correspond to three levels of polygenic overlap: perfect polygenic overlap ($\pi_{12} = \pi_1^u$), partial polygenic overlap equal to 10% of the polygenicity of the traits ($\pi_{12} = \pi_1^u/10$), and random polygenic overlap, which arise by chance if markers are spread randomly throughout the genome ($\pi_{12} = \pi_1^u \pi_2^u$). The simulations were performed with complete overlap of GWAS samples (N = 100,000), heritability h^2 of 0.5, and equal polygenicity in both traits ($\pi = \pi_1^u = \pi_2^u$) ranging from 10^{-3} (high polygenicity), 10^{-4} (medium polygenicity) to 10^{-5} (low polygenicity). The average of 10 simulations is shown with averaged estimated standard errors.

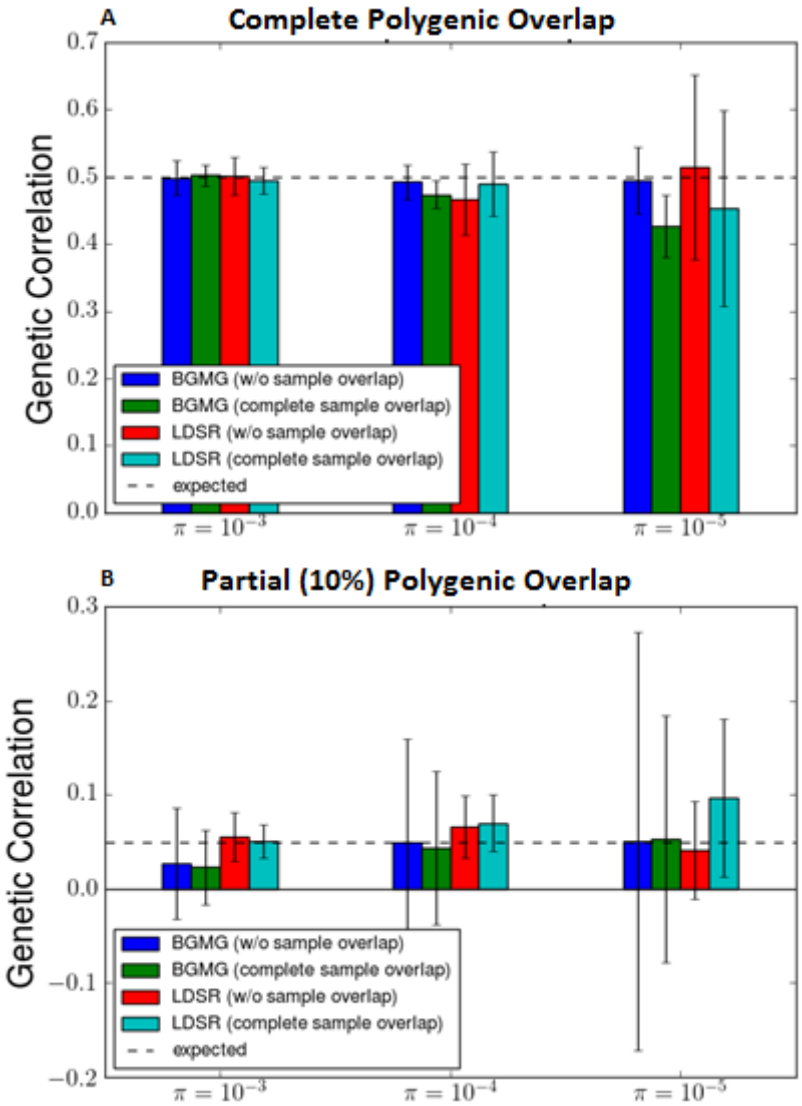

Figure 4: Estimates of genetic correlation from Bivariate Mixture Model for GWAS (BGMG) and Cross-trait LD Score Regression (LDSR), with complete sample overlap (N = 100,000) and without sample overlap (N = 50,000 in each sample) between GWAS studies. In all simulations ρ_{12} is fixed at 0.5. Expected genome-wide genetic correlation is r_g = 0.5 for complete polygenic overlap, and r_g = 0.05 for partial polygenic overlap.

REFERENCES

1. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
2. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607-18 (2011).
3. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483-95 (2013).
4. Schork, A.J., Wang, Y., Thompson, W.K., Dale, A.M. & Andreassen, O.A. New statistical approaches exploit the polygenic architecture of schizophrenia--implications for the underlying neurobiology. *Curr Opin Neurobiol* **36**, 89-98 (2016).
5. Pasaniuc, B. & Price, A.L. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet* **18**, 117-127 (2017).
6. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-94 (2013).
7. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
8. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
9. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89-98 (2014).
10. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
11. Smeland, O.B. *et al.* Identification of Genetic Loci Jointly Influencing Schizophrenia Risk and the Cognitive Traits of Verbal-Numerical Reasoning, Reaction Time, and General Cognitive Function. *JAMA Psychiatry* (2017).
12. Hagenaars, S.P. *et al.* Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Mol Psychiatry* **21**, 1624-1632 (2016).
13. Hill, W.D. *et al.* Age-Dependent Pleiotropy Between General Cognitive Function and Major Psychiatric Disorders. *Biol Psychiatry* **80**, 266-73 (2016).
14. Holland, D. *et al.* Estimating Degree Of Polygenicity, Causal Effect Size Variance, And Confounding Bias In GWAS Summary Statistics. *bioRxiv* (2017).
15. Gottesman, II & Gould, T.D. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**, 636-45 (2003).
16. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
17. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186.
18. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304-5 (2011).
19. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
20. Smeland, O.B. *et al.* Genetic overlap between schizophrenia and volumes of hippocampus, putamen and intracranial volume indicates shared molecular genetic mechanisms. *Schizophrenia Bulletin (in publication)* (2017).
21. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**, 709-17 (2016).

22. Holland, D. *et al.* Estimating inflation in GWAS summary statistics due to variance distortion from cryptic relatedness. *bioRxiv* (2017).
23. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* **42**, 570-5 (2010).
24. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genetics* **9**, e1003348 (2013).
25. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-8 (2015).
26. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-45 (2012).
27. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
28. Turley, P. *et al.* MTAG: Multi-Trait Analysis of GWAS. *bioRxiv* (2017).
29. Hu, Y. *et al.* Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet* **13**, e1006836 (2017).
30. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput Biol* **13**, e1005589 (2017).
31. Vilhjalmsson, B.J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576-92 (2015).
32. Schork, A.J. *et al.* All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet* **9**, e1003449 (2013).

ONLINE METHODS

This article is accompanied by a Supplementary Note with further details.

Model for bivariate distribution of GWAS z-scores. We use method of moments to derive the following model for GWAS z-scores (see Supplementary note):

$$\left(z_{1j}, z_{2j}\right) = \left(\delta_{1j}, \delta_{2j}\right) + N\left((0,0), \begin{bmatrix} \sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02} \\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2 \end{bmatrix}\right) \tag{4}$$

$$\left(\delta_{1j}/\sqrt{N_{1j}}, \delta_{2j}/\sqrt{N_{2j}}\right) \sim \pi'_{0j}N(0,0) + \pi'_{1j}N\left(0, \Sigma'_{1j}\right) + \\ + \pi'_{2j}N\left(0, \Sigma'_{2j}\right) + \pi'_{12,j}N\left(0, \Sigma'_{12,j}\right),$$

where $\pi'_{cj} = \ell_j\pi_c/\eta_j$ is SNP-adjusted weight of mixture component $c \in 1,2,12$; $\Sigma'_{cj} = \eta_j\Sigma_c$ is SNP-adjusted variance-covariance matrix; $\ell_j = \sum_i H_i r_{ij}^2$ is the LD score (adjusted for heterozygosity[33]); and $\eta_{cj} = \left(\pi_c\ell_j + (1-\pi_c)\frac{\sum_i H_i^2 r_{ij}^4}{\sum_i H_i r_{ij}^2}\right)$ can be interpreted as shape parameter that affects fourth and higher moments of the distribution. This model explains second moments $E[Z_{1j}^2]$, $E[Z_{1j}Z_{2j}]$, $E[Z_{2j}^2]$ and fourth moments $E[Z_{1j}^4]$, $E[Z_{1j}^2Z_{2j}^2]$, $E[Z_{2j}^4]$ of z score distribution, and is consistent with a mixture model of sparse and ubiquitous effects[34,35]. Note that the model involves the forth power of allelic correlation r_{ij}^4, which is directly proportional to kurtosis (measure of heavy tails) of z-score distribution.

LD score estimation. To estimate LD scores, we follow the procedure from LD score regression method[10,16,27]. For r^2, we calculate unbiased estimate[36] of r^2 across 1 centimorgan (cm) window without cutoff for small r^2 value. For r^4, we lacked analytical expression for unbiased estimate, and instead calculated the ratio $\sum_i H_i^2 r_{ij}^4 \,/\, \sum_i H_i r_{ij}^2$ across biased estimates of r^2 and r^4 with cutoff $r^2 \geq 0.05$. For simulations LD scores were estimated from the genotypes that we use to produce synthetic GWAS data.

Fit procedure. We fit the model by direct optimization of weighted log likelihood

$$F(\theta) = \sum_j w_j \log(pdf(z_j|\theta)), \tag{5}$$

where $\theta = (\pi_1, \pi_2, \pi_{12}, \sigma_1^2, \sigma_2^2, \rho_{12}, \sigma_{01}^2, \sigma_{02}^2, \rho_0)$ is a vector of all parameters being optimized, and weights w_j chosen inversely proportional to the LD score. Optimization is done by Nelder-Mead Simplex Method[37] as implemented in MATLAB's `fminsearch`. First, we fit univariate parameters separately for each trait (i.e. $\pi_1^u, \sigma_1^2, \sigma_{01}^2$ for the first trait, and similarly for the second trait). We employ a sequence of optimizations to ensure robust convergence. First, we use infinitesimal model $\pi_1^u = 1$ to find $\sigma_{1,inf}^2$ and to initialize σ_{01}^2; second, we use constraint $\pi_1^u\sigma_1^2 = \sigma_{1,inf}^2$ to find initial values of π_1^u and σ_1^2. Third, we use unconstrained optimization to jointly optimize $\pi_1^u, \sigma_1^2, \sigma_{01}^2$, and repeat the same procedure to find $\pi_2^u, \sigma_2^2, \sigma_{02}^2$. In bivariate optimization we again use infinitesimal model $\pi_{12} = 1$ to initialize ρ_{12} and ρ_0, and then proceed with unconstrained optimization of all parameters.

Standard error estimation. We estimate standard errors of all parameters from observed Fisher's information, which is the standard method in likelihood optimization

theory. The limitation of this method is that it is not suitable for parameters near their boundary, which is especially applicable to mixture weights π_1, π_2 and π_{12}. To avoid this problem we apply transformations — MATLAB's `logit()` for π_1, π_2, π_{12}, `exp()` for $\sigma_1^2, \sigma_2^2, \sigma_{01}^2, \sigma_{02}^2$, and `erf()` for ρ_0, ρ_{12}, and estimated variance-covariance matrix of errors in the transformed parameter space. We validated that our estimates based on observed Fisher's information are in good agreement with block jack knife estimates. To estimate standard errors for functions of the parameters, such as r_g and h^2, we incorporate linear correlation among parameter errors in transformed space. We sample N=1000 realizations of the parameter vector, calculating the function (e.g., r_g or h^2) on each of them, and report the 95% confidence interval and standard errors.

Genetic correlation. Parameter ρ_{12} in BGMG model plays the role of genetic correlation, calculated from a subset of variants affecting both traits. Genome-wide genetic correlation, calculated across all SNPs, is related to ρ_{12} by the following formula that involves polygenicity π_1^u and π_2^u of the traits, and polygenic overlap π_{12}:

$$r_g = \rho_{12}\pi_{12}/\sqrt{\pi_1^u \pi_2^u} \tag{6}$$

For traits with K-fold difference in polygenicity ($\pi_1^u = K\pi_2^u$) this formula predicts an upper bound on genome-wide genetic correlation: $r_g \leq \rho_{12}/\sqrt{k}$, where equality holds if causal variants of the less polygenic trait form a subset of the higher-polygenic trait markers.

Large LD blocks. We use inverse LD score weighting to avoid overcounting evidence from large LD blocks. An alternative approach, also available in the BGMG implementation, is to perform random pruning – a stochastic procedure that average log likelihood function across repeatedly selected subsets J of variants such that for each pair of variants $i, j \in J$ the squared allelic correlation r_{ij}^2 falls below certain threshold. Given T iterations of random pruning the log-likelihood function can be calculated as follows:

$$F(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j\in J_t}\log(pdf(z_j|\theta)) \tag{7}$$

which is equivalent to weighted log-likelihood $F(\theta) = \sum_j w_j \log(pdf(z_j|\theta))$ with weights $w_j = |\{t: j \in J_t\}|/T$, $|S|$ denotes cardinality of set S. We refer to this as "random-pruning induced weights". Empirically, estimates based on inverse LD weighting are consistent with estimates based on random-pruning induced weights with cutoff $r^2 = 0.1$.

SNPs in the analysis. To enable direct comparison of our model with LD score regression we use the same set of SNPs in our log likelihood optimization, which consist of approx. 1.1 million variants, subset of 1000 Genomes and HapMap3[38], with MAF >= 0.05, ambiguous SNPs excluded, imputation INFO above 0.9, MHC and other long-range LD regions excluded. This set of SNPs is also used to calculate weights in log-likelihood optimization. Calculation of the LD scores ℓ_j and shape parameter η_j are based on the approx. 10 million SNPs, available from 1000 Genomes Phase 3 data.

Simulations. We generated genotypes for 10^5 unrelated simulated samples using HapGen2[18]. To generate a quantitative phenotype y_k of k-th sample we use simple additive genetic model, $y_k = \sum_j g_{kj}\beta_j + \epsilon_k$, where g_{kj} is the number of reference alleles for j-th SNP on k-th sample, β_j is causal effect size drawn according to bivariate model (1), and ϵ is the residual vector drawn from normal distribution with zero mean and variance chosen in a way that sets heritability $h^2 = var(\boldsymbol{G}\beta)/var(y)$ to a predefined level. To generate GWAS p-values we note that the regression slope, $\hat{\beta}_j$, and the Pearson correlation coefficient, $r_j = corr(y, g_{\cdot j})$, are assumed to be t-distributed. These quantities have the same t-value: $t_j = \beta_j/se(\beta_j) = r_j/se(r_j) = r_j\sqrt{N-2}/\sqrt{1-r_j^2}$, and therefore the same p-value, equal to Student's t cumulative distribution function (cdf) with $N-2$ degrees of freedom: $P_{val,j} = 2\, tcdf(-|t_j|, N-2)$, where N is the sample size. Since we are not here dealing with covariates, we calculated p-value from correlation r_j, which is slightly faster than from estimating the regression coefficient.

METHODS-ONLY REFERENCES

33. Thompson, W.K. *et al.* An Empirical Bayes Mixture Model for Effect Size Distributions in Genome-Wide Association Studies. *PLoS Genet* **11**, e1005717 (2015)
34. Holland, D. *et al.* Estimating Effect Sizes and Expected Replication Probabilities from GWAS Summary Statistics. *Frontiers in Genetics* **7**, 15 (2016).
35. Wang, Y. *et al.* Leveraging Genomic Annotations and Pleiotropic Enrichment for Improved Replication Rates in Schizophrenia GWAS. *PLOS Genetics* **12**, e1005803 (2016).
36. Wherry, R.J. A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation. *Ann. Math. Statist.* **2**, 440-457 (1931).
37. Lagarias, J.C., Reeds, J.A., Wright, M.H. & Wright, P.E. Convergence properties of the Nelder--Mead simplex method in low dimensions. *SIAM Journal on optimization* **9**, 112-147 (1998).
38. International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).

SUPPLEMENTARY NOTE

Separate PDF document.